Topic 2: Summary Measure For Populations

▲Recall, parameters are characteristics of _______.
▲Recall, parameters summarize data.

We need to be able to compute such measures for individual and "grouped" data.

"The most commonly used parameters for interpreting and understanding the meaning of values in populations are measures of ________tendency and variability. They summarize data for logical presentation."

Measures of Location

☐There are several measures of "_________ **tendency**":

(i) The **M_____**: The value that occurs most often. Or in a frequency distribution, it is the point or class mark corresponding to the _____ with the highest frequency.

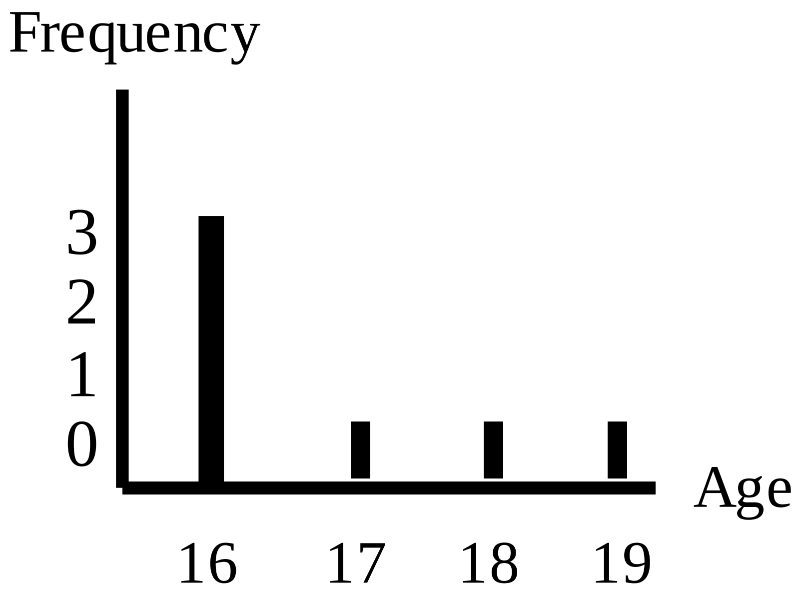
Example: Age of students in driving school:

{16, 16, 17, 18, 16, 19.}

Mode=___

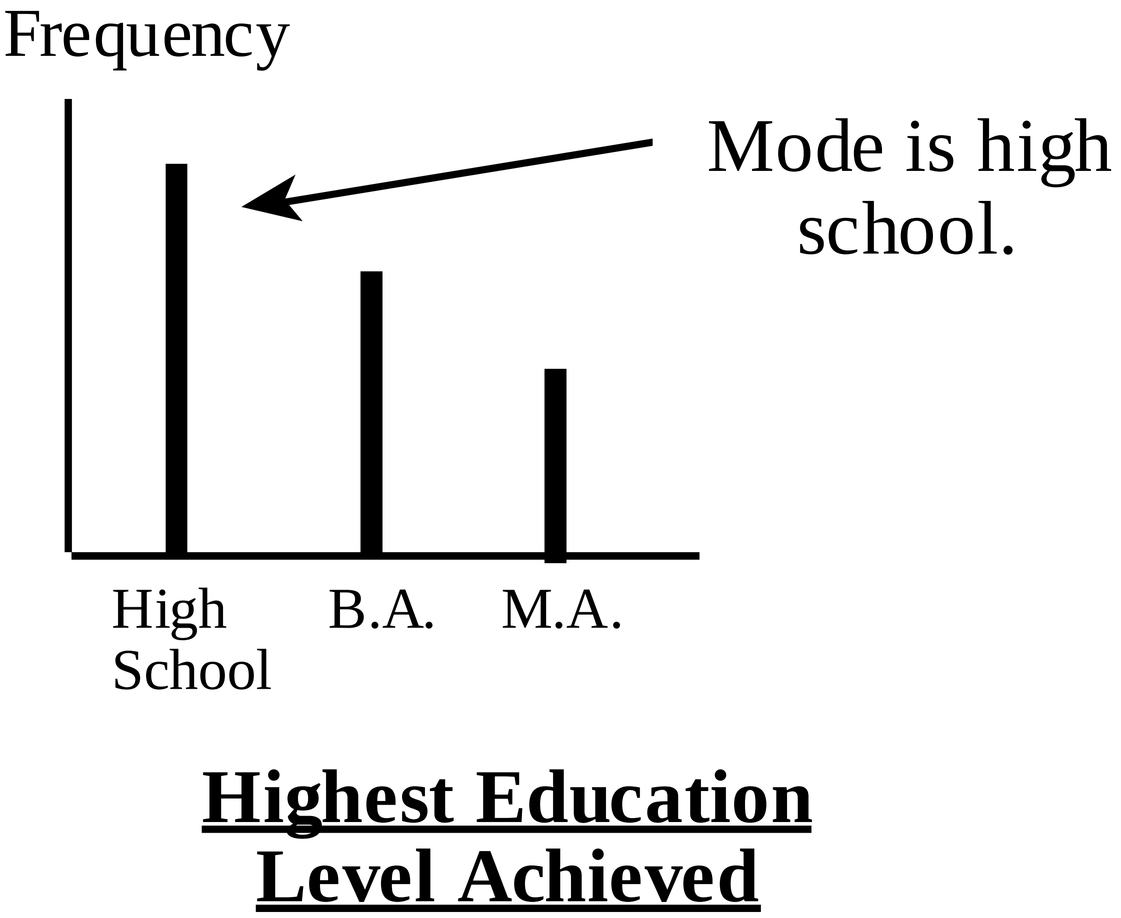
<u>Problems with Mode</u> (as a measure of central tendency):

(1) Mode may not be near the_______of the data.
(2) Data may have <u>more</u> than one ______.



Mode is not representative of central location of the distribution of ages.

However, for purely **descriptive** purposes, the ______ can be useful in representing the ______ frequently occurring value:



Mode is high school.

Example: Using the mode with a **frequency distribution**:

Income	**Frequency**
80 $\leq$ X < 100	
100 $\leq$ X < 120	
120 $\leq$ X < 140	
140 $\leq$ X < 160	**6**
160 $\leq$ X < 180	**3**
180 $\leq$ X < 200	***13***
200 **$\leq$ X < 220**	***2***
	$\sum f_i = 40$

Class containing the most observations.

- Modal Interval is **{180 $\leq$ X < 200}.**

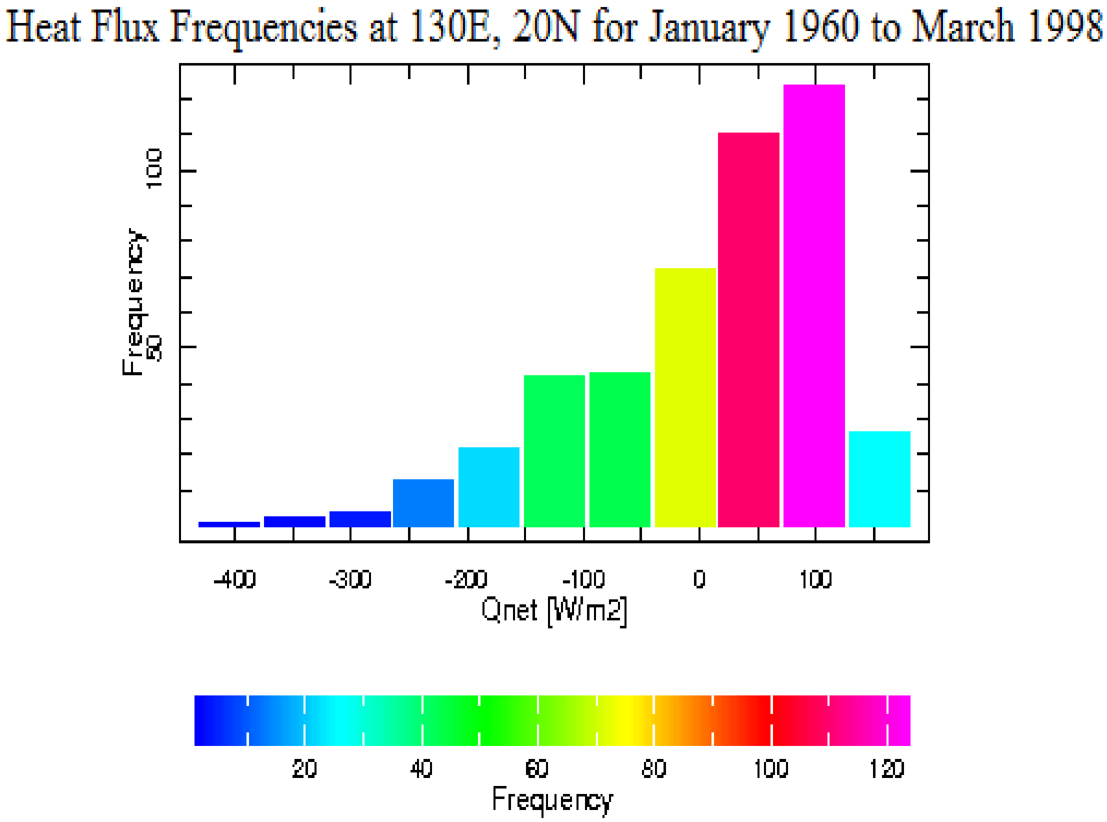
- This is the mode because it is the interval with the ___________________ within it.

- 190 is also considered the "_______" because it is the mid-point of **{180 $\leq$ X < 200}**.

Both are correct!

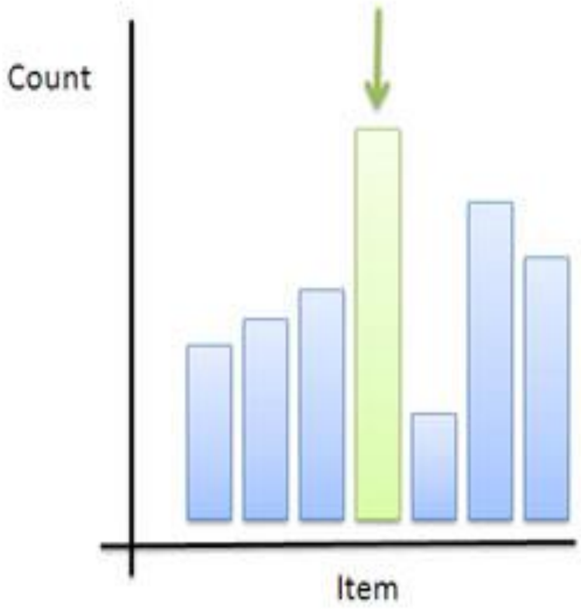
The Mode:

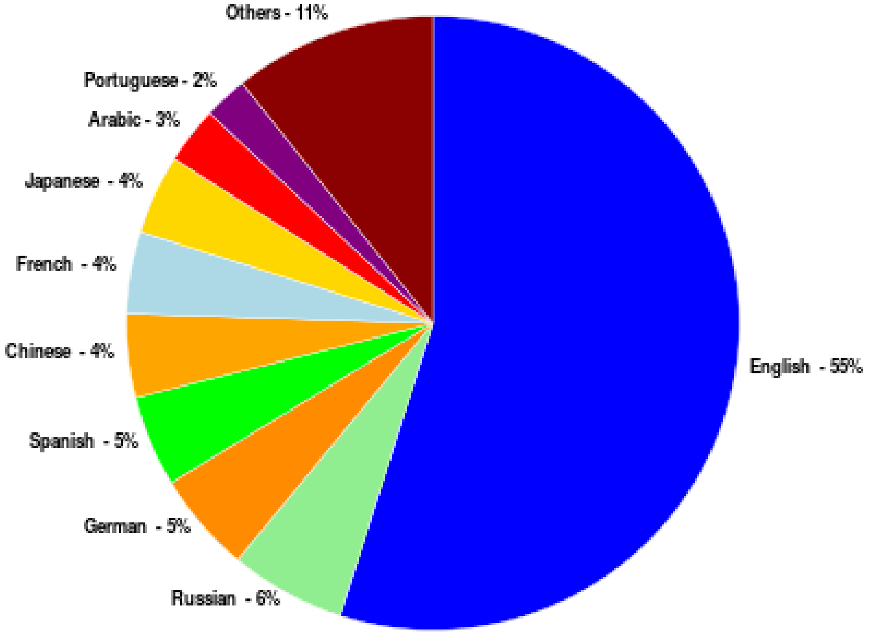
The mode is defined as the most frequently observed value. For grouped data, the mode is the most commonly observed category, and for ungrouped data, the mode is the value which occurs most frequently.



The histogram is unimodal and is negatively skewed.

Mode (Most Popular)





April 2013

"Another measure of central tendency is the Median":

(ii) The **M______**: The m_________value in a set of numbers arranged in order of magnitude.

I.e. the "middle" ranked value in the data.

(Put into ____________order first!)

Example: Sales: N=9 (odd number of values)

{14, 15, 18, 19, **20**, 21, 23, 25, 26}

Median

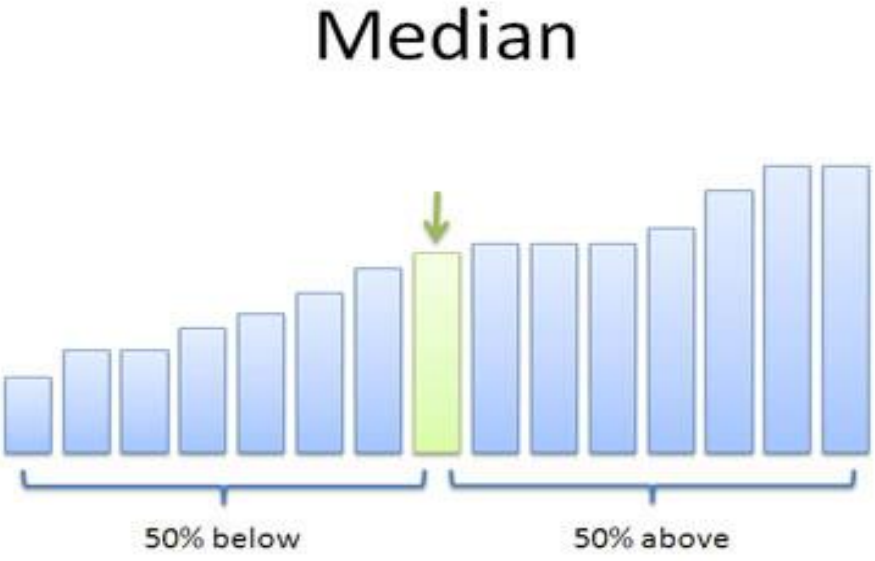
Example: Sales: N=12 (even number of values)

$$\{16, 16, 17, 18, 19, \underbrace{\mathbf{20, 21}}_{\text{median} = 20.5}, 22, 23, 24, 25, 27\}$$

6th

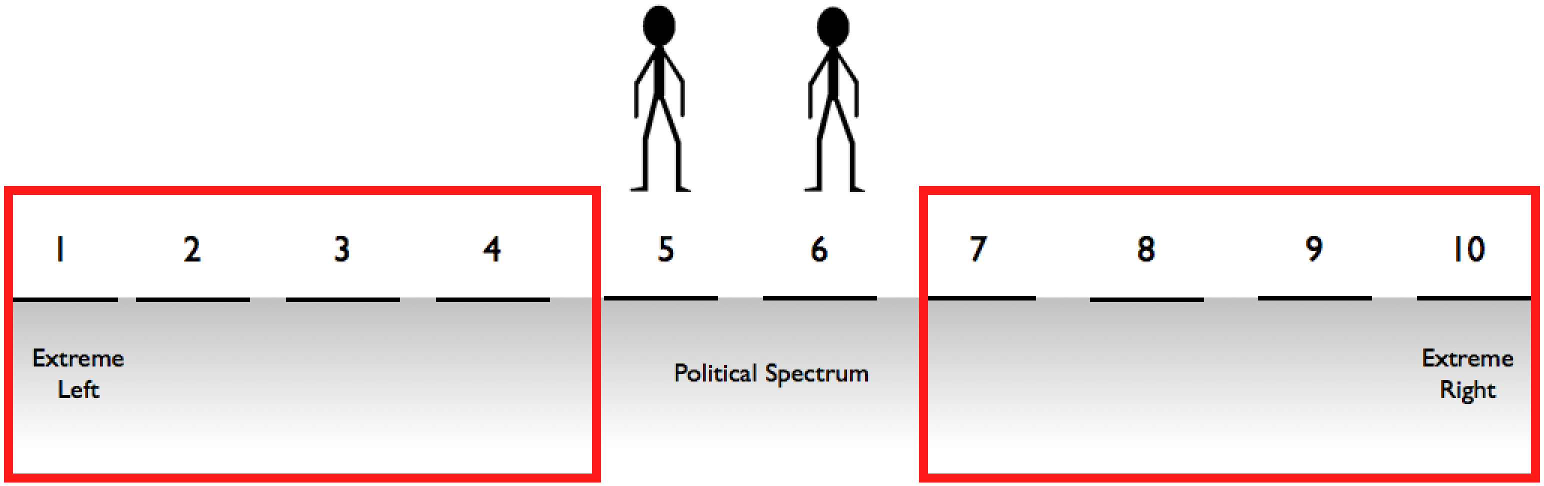
7th

Note: $\frac{20+21}{2} = 20.5$



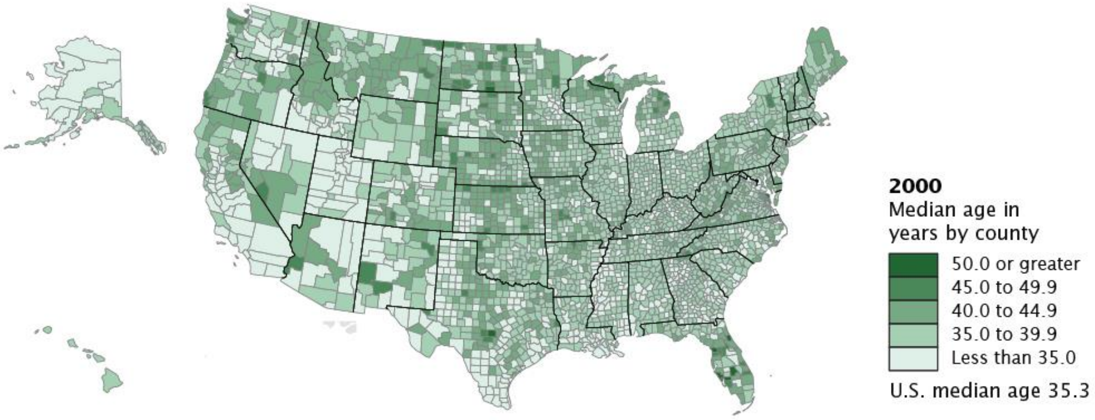
The Median Voter Theorem

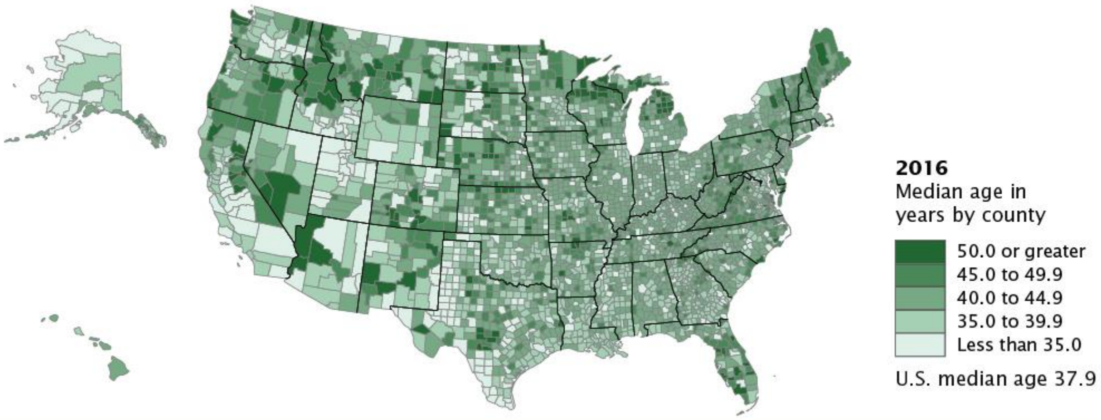
Candidates position themselves in the middle of the spectrum



Locations 1-4 and 7-10 are elminated *via* the iterated deletion of dominated strategies

The Nation's Median Age Continues to Rise





U.S. Department of Commerce
Economics and Statistics Administration
U.S. CENSUS BUREAU
census.gov

Sources: Census 2000 Summary File 1 and Vintage 2016 Population Estimates
www.census.gov/census2000/sumfile1.html
www.census.gov/programs-surveys/popest.html

Example: Using the median with a frequency distribution:

Income	**Frequency**	**Cumulative Frequency**
80 $\leq$ X < 100		**2**
100 $\leq$ X < 120		**8**
120 $\leq$ X < 140		**16**
140 $\leq$ X < 160		**22**
160 $\leq$ X < 180		**25**
180 $\leq$ X < 200	**13**	**38**
200 $\leq$ X < 220	**2**	**40=N**
	$\sum f_i = 40$	

N=40

40\2=20 which is an even number.

Find the interval that contains the 20th and 21st ranked observations.

Median is half-way between the 20th and 21st ranked observations –

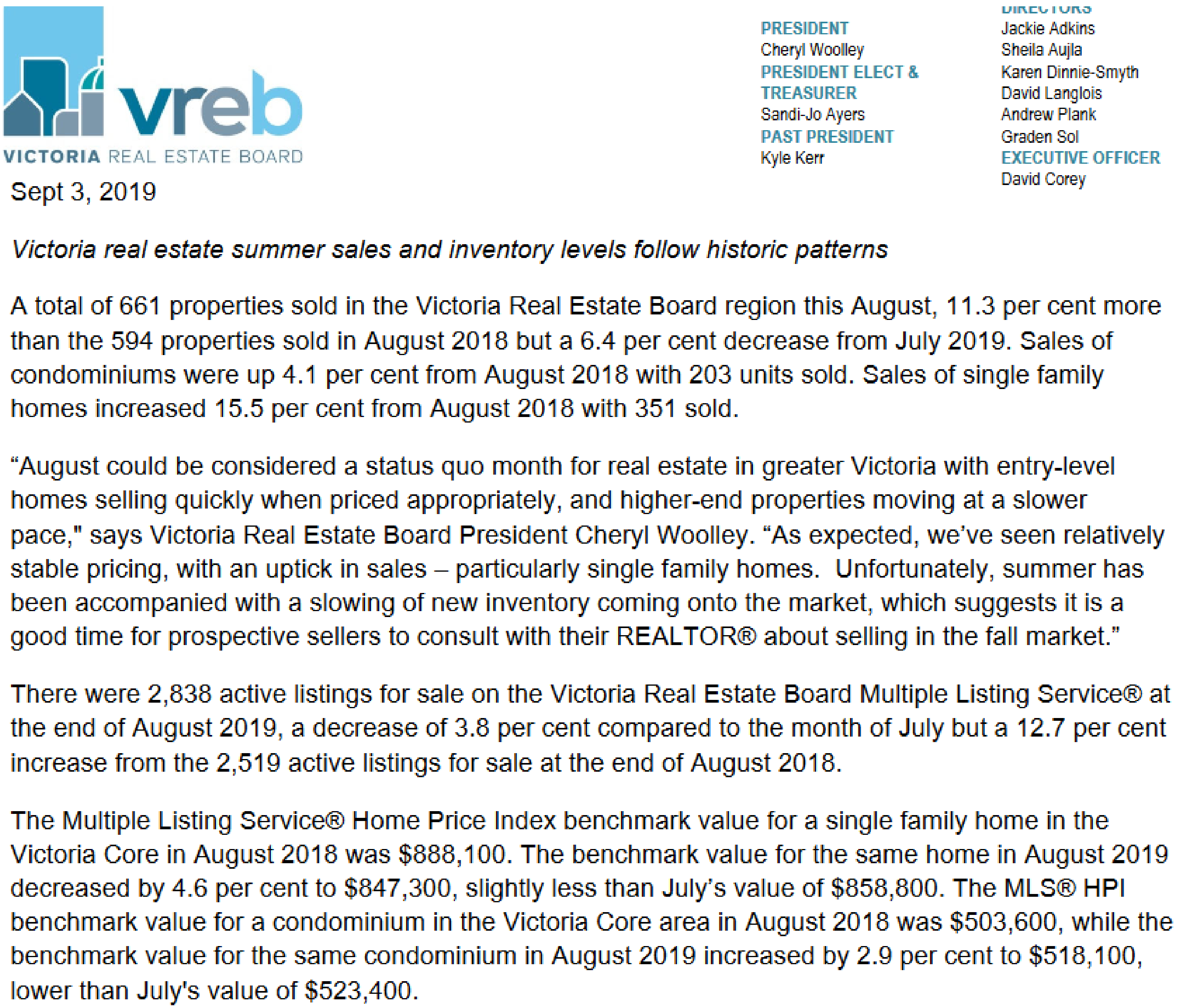
⇨**Median class is (___ to ____).**

Example: 20th observation on the boundary or 20th and 21st observations in different intervals:

Income	**Frequency**	**Cumulative Frequency**
80 $\leq$X < 100		**2**
100 $\leq$X < 120		**14**
120 $\leq$X < 140		**20**
140 $\leq$X < 160		**24**
160 $\leq$X < 180		**34**
180 $\leq$X < 200		**37**
200 $\leq$X < 220		**40=N**
	$\sum f_i = 40$	

If we assumed that the 20th and 21st observations are in different intervals:

Median 120 to 160 **or** Median =140.



VICTORIA REAL ESTATE BOARD

Sept 3, 2019

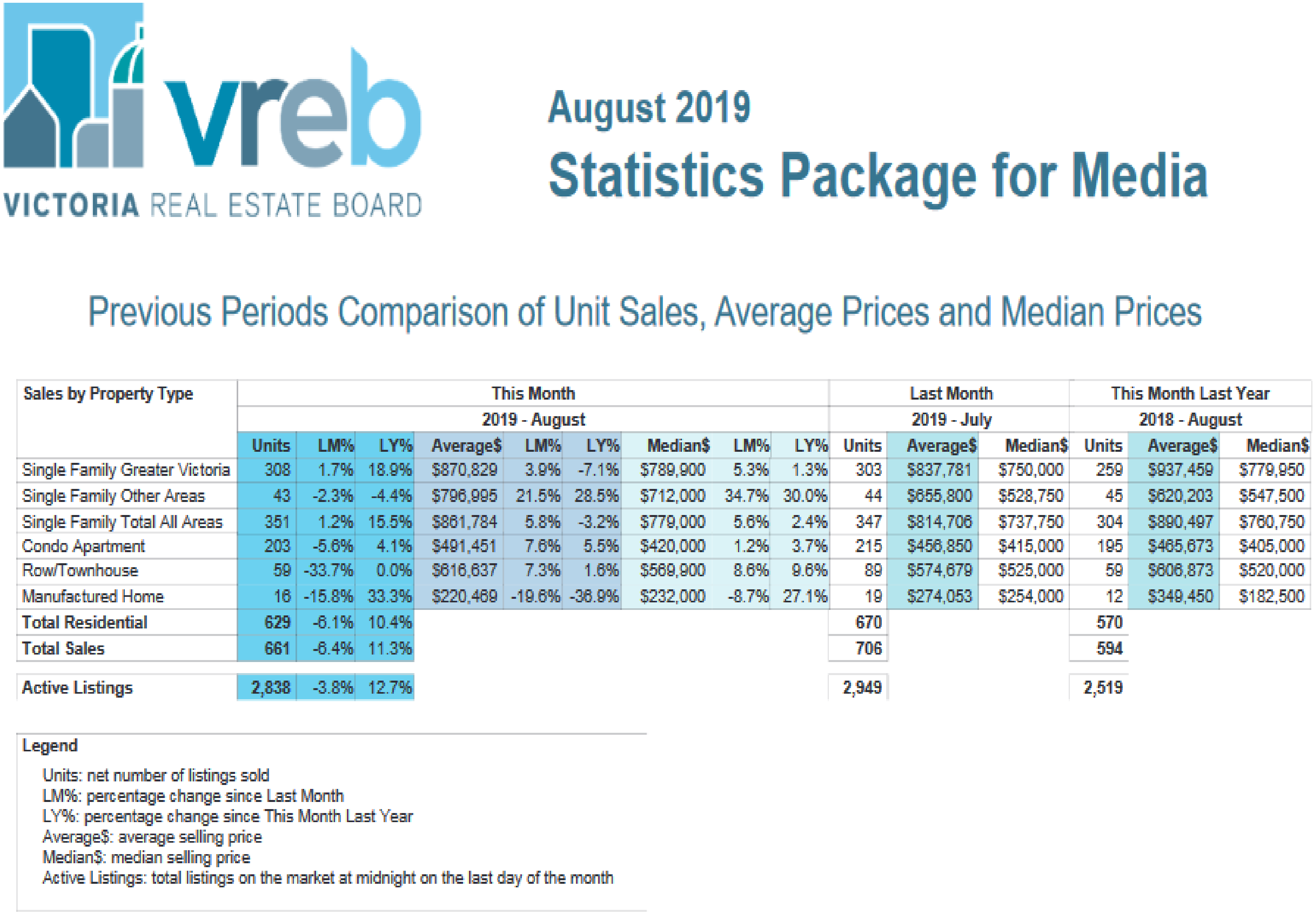
Victoria real estate summer sales and inventory levels follow historic patterns

A total of 661 properties sold in the Victoria Real Estate Board region this August, 11.3 per cent more than the 594 properties sold in August 2018 but a 6.4 per cent decrease from July 2019. Sales of condominiums were up 4.1 per cent from August 2018 with 203 units sold. Sales of single family homes increased 15.5 per cent from August 2018 with 351 sold.

"August could be considered a status quo month for real estate in greater Victoria with entry-level homes selling quickly when priced appropriately, and higher-end properties moving at a slower pace," says Victoria Real Estate Board President Cheryl Woolley. "As expected, we've seen relatively stable pricing, with an uptick in sales – particularly single family homes. Unfortunately, summer has been accompanied with a slowing of new inventory coming onto the market, which suggests it is a good time for prospective sellers to consult with their REALTOR® about selling in the fall market."

There were 2,838 active listings for sale on the Victoria Real Estate Board Multiple Listing Service® at the end of August 2019, a decrease of 3.8 per cent compared to the month of July but a 12.7 per cent increase from the 2,519 active listings for sale at the end of August 2018.

The Multiple Listing Service® Home Price Index benchmark value for a single family home in the Victoria Core in August 2018 was $888,100. The benchmark value for the same home in August 2019 decreased by 4.6 per cent to $847,300, slightly less than July's value of $858,800. The MLS® HPI benchmark value for a condominium in the Victoria Core area in August 2018 was $503,600, while the benchmark value for the same condominium in August 2019 increased by 2.9 per cent to $518,100, lower than July's value of $523,400.



vreb
VICTORIA REAL ESTATE BOARD

August 2019

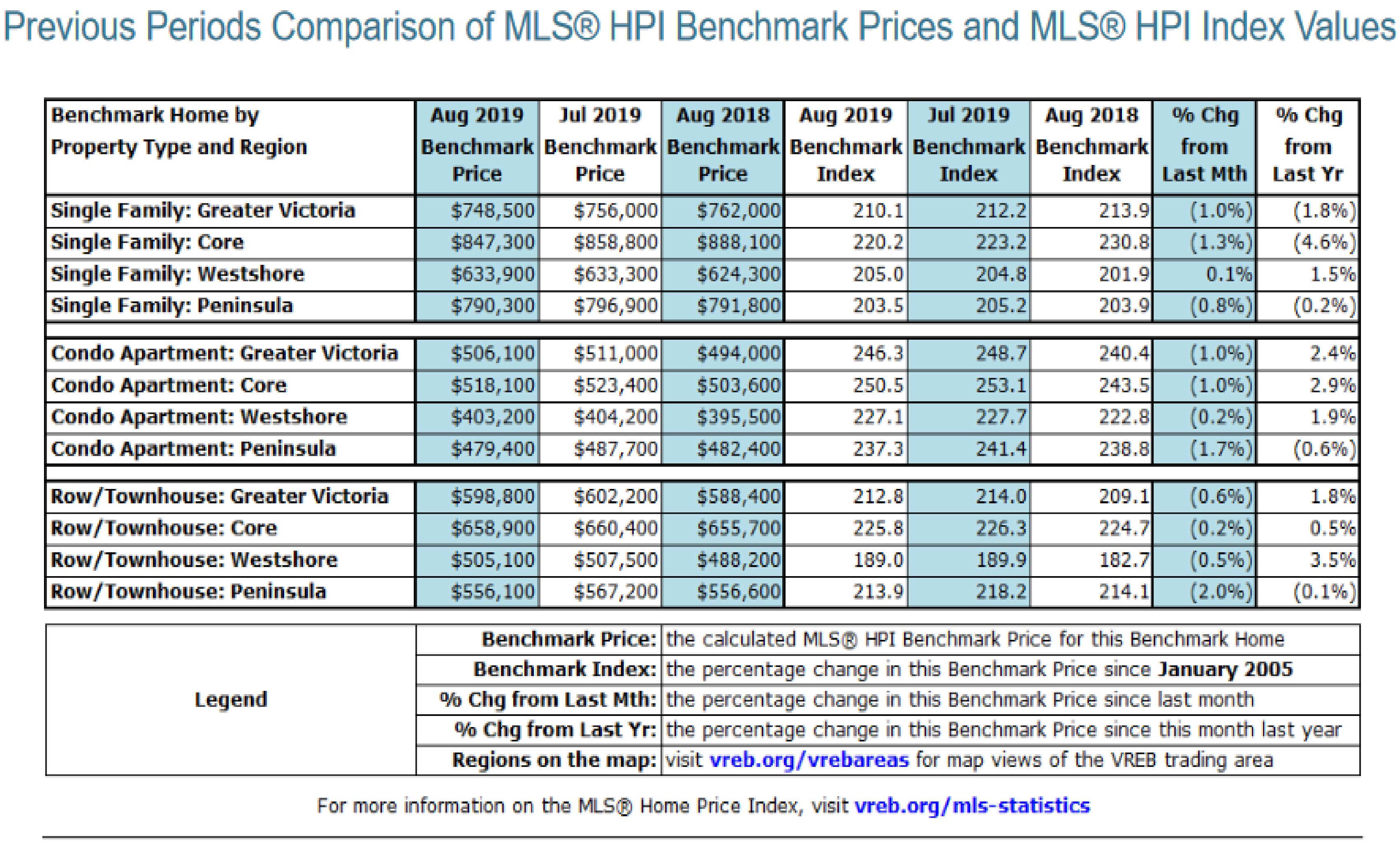
Statistics Package for Media

Previous Periods Comparison of Unit Sales, Average Prices and Median Prices

Sales by Property Type	This Month									Last Month			This Month Last Year		
	2019 - August									2019 - July			2018 - August		
	Units	LM%	LY%	Average$	LM%	LY%	Median$	LM%	LY%	Units	Average$	Median$	Units	Average$	Median$
Single Family Greater Victoria	308	1.7%	18.9%	$870,829	3.9%	-7.1%	$789,900	5.3%	1.3%	303	$837,781	$750,000	259	$937,459	$779,950
Single Family Other Areas	43	-2.3%	-4.4%	$796,995	21.5%	28.5%	$712,000	34.7%	30.0%	44	$655,800	$528,750	45	$620,203	$547,500
Single Family Total All Areas	351	1.2%	15.5%	$861,784	5.8%	-3.2%	$779,000	5.6%	2.4%	347	$814,706	$737,750	304	$890,497	$760,750
Condo Apartment	203	-5.6%	4.1%	$491,451	7.6%	5.5%	$420,000	1.2%	3.7%	215	$456,850	$415,000	195	$465,673	$405,000
Row/Townhouse	59	-33.7%	0.0%	$616,637	7.3%	1.6%	$569,900	8.6%	9.6%	89	$574,679	$525,000	59	$606,873	$520,000
Manufactured Home	16	-15.8%	33.3%	$220,469	-19.6%	-36.9%	$232,000	-8.7%	27.1%	19	$274,053	$254,000	12	$349,450	$182,500
Total Residential	**629**	-6.1%	10.4%							**670**			**570**		
Total Sales	**661**	-6.4%	11.3%							**706**			**594**		
Active Listings	**2,838**	-3.8%	12.7%							**2,949**			**2,519**		

Legend

Units: net number of listings sold
LM%: percentage change since Last Month
LY%: percentage change since This Month Last Year
Average$: average selling price
Median$: median selling price
Active Listings: total listings on the market at midnight on the last day of the month



Previous Periods Comparison of MLS® HPI Benchmark Prices and MLS® HPI Index Values

Benchmark Home by Property Type and Region	Aug 2019 Benchmark Price	Jul 2019 Benchmark Price	Aug 2018 Benchmark Price	Aug 2019 Benchmark Index	Jul 2019 Benchmark Index	Aug 2018 Benchmark Index	% Chg from Last Mth	% Chg from Last Yr
Single Family: Greater Victoria	$748,500	$756,000	$762,000	210.1	212.2	213.9	(1.0%)	(1.8%)
Single Family: Core	$847,300	$858,800	$888,100	220.2	223.2	230.8	(1.3%)	(4.6%)
Single Family: Westshore	$633,900	$633,300	$624,300	205.0	204.8	201.9	0.1%	1.5%
Single Family: Peninsula	$790,300	$796,900	$791,800	203.5	205.2	203.9	(0.8%)	(0.2%)
Condo Apartment: Greater Victoria	$506,100	$511,000	$494,000	246.3	248.7	240.4	(1.0%)	2.4%
Condo Apartment: Core	$518,100	$523,400	$503,600	250.5	253.1	243.5	(1.0%)	2.9%
Condo Apartment: Westshore	$403,200	$404,200	$395,500	227.1	227.7	222.8	(0.2%)	1.9%
Condo Apartment: Peninsula	$479,400	$487,700	$482,400	237.3	241.4	238.8	(1.7%)	(0.6%)
Row/Townhouse: Greater Victoria	$598,800	$602,200	$588,400	212.8	214.0	209.1	(0.6%)	1.8%
Row/Townhouse: Core	$658,900	$660,400	$655,700	225.8	226.3	224.7	(0.2%)	0.5%
Row/Townhouse: Westshore	$505,100	$507,500	$488,200	189.0	189.9	182.7	(0.5%)	3.5%
Row/Townhouse: Peninsula	$556,100	$567,200	$556,600	213.9	218.2	214.1	(2.0%)	(0.1%)

Legend	**Benchmark Price:**	the calculated MLS® HPI Benchmark Price for this Benchmark Home
	Benchmark Index:	the percentage change in this Benchmark Price since **January 2005**
	% Chg from Last Mth:	the percentage change in this Benchmark Price since last month
	% Chg from Last Yr:	the percentage change in this Benchmark Price since this month last year
	Regions on the map:	visit **vreb.org/vrebareas** for map views of the VREB trading area

For more information on the MLS® Home Price Index, visit **vreb.org/mls-statistics**

MLS®
Home Price Index

HPI or Benchmark Price
○ HPI
● Benchmark Price

Value or percent change
● Value
○ Percent change

1. Area Group
VREB District Summary

2. Property Type
Single Family-All (SF-All)

3. Area/Property Type Selection
All

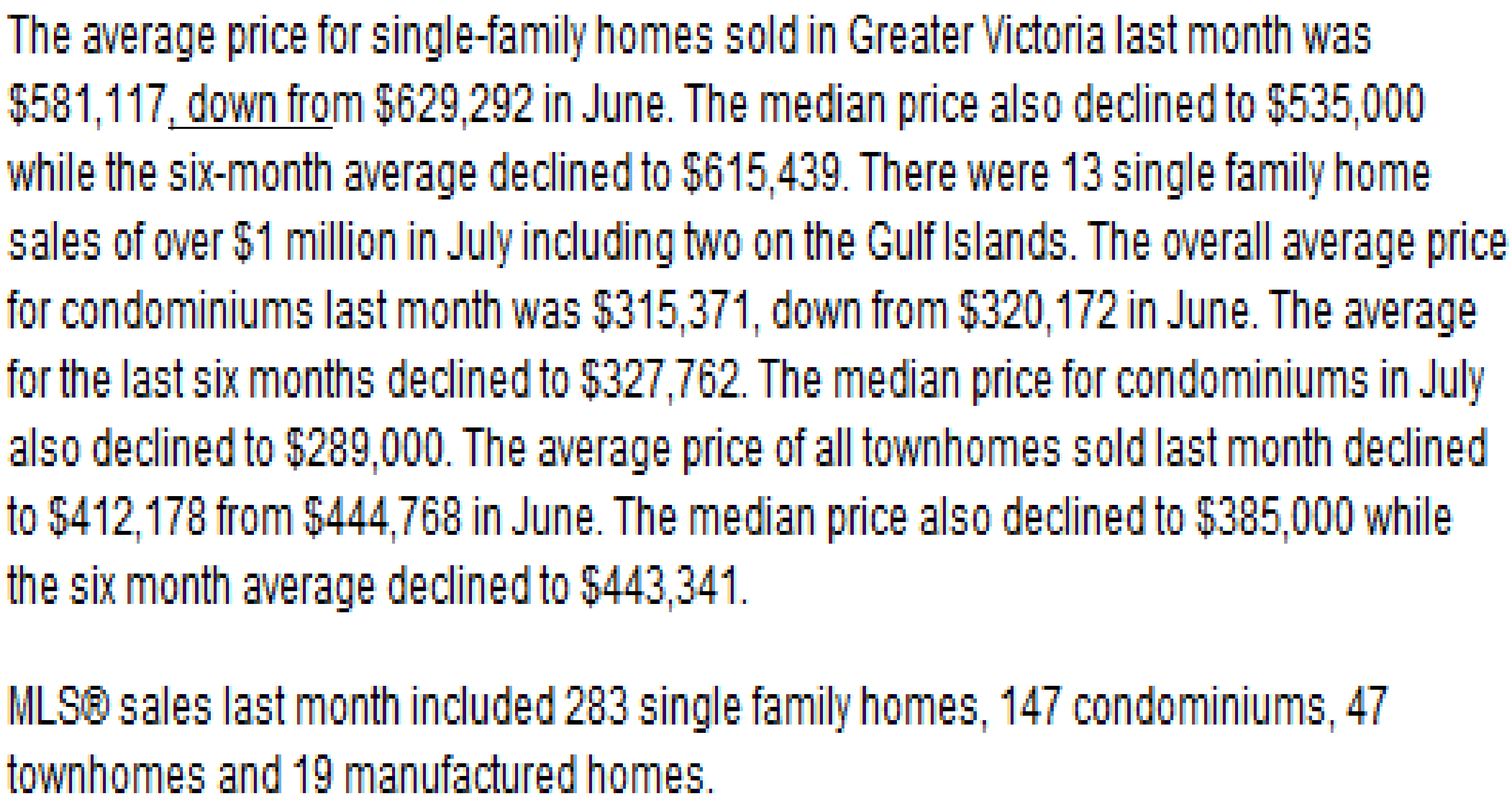
Benchmark Price by Timeframe and Property Type

	August 2019	1 Month Ago	3 Months Ago	6 Months Ago	12 Months Ago	3 Years Ago	5 Years Ago	January 2005
Victoria REB – SF-All	$748,500	$756,000	$756,000	$735,300	$762,000	$635,900	$500,500	$356,300
Victoria – SF-All	$818,700	$831,700	$835,000	$813,500	$856,200	$736,400	$536,200	$370,800
Victoria West – SF-All	$664,100	$673,200	$674,700	$634,900	$701,400	$560,700	$427,700	$284,200
Oak Bay – SF-All	$1,197,700	$1,212,600	$1,203,800	$1,181,200	$1,310,900	$1,108,900	$760,600	$552,000
Esquimalt – SF-All	$638,000	$657,300	$653,200	$622,200	$679,800	$551,400	$438,900	$292,400
View Royal – SF-All	$754,800	$765,600	$753,800	$718,300	$778,100	$622,200	$507,600	$347,200
Saanich East – SF-All	$860,000	$871,100	$877,400	$860,400	$902,600	$774,000	$576,800	$394,500
Saanich West – SF-All	$747,800	$754,400	$756,000	$736,900	$758,700	$620,400	$486,200	$329,900
Sooke – SF-All	$536,900	$543,300	$548,900	$529,500	$514,700	$400,800	$353,900	$264,900
Langford – SF-All	$638,800	$637,500	$641,500	$620,000	$634,800	$500,200	$429,300	$308,000
Metchosin – SF-All	$920,400	$933,100	$946,300	$923,100	$954,100	$724,500	$621,200	$455,400
Colwood – SF-All	$681,900	$665,600	$666,600	$660,600	$676,600	$548,800	$451,400	$332,600
Highlands – SF-All	$832,000	$838,600	$846,000	$864,800	$842,100	$714,700	$595,200	$437,700
North Saanich – SF-All	$946,900	$958,100	$940,600	$934,700	$947,900	$805,600	$617,000	$487,300
Sidney – SF-All	$670,300	$679,600	$660,600	$646,100	$660,000	$569,000	$435,900	$321,500

Monthly Sales Summary July 2011

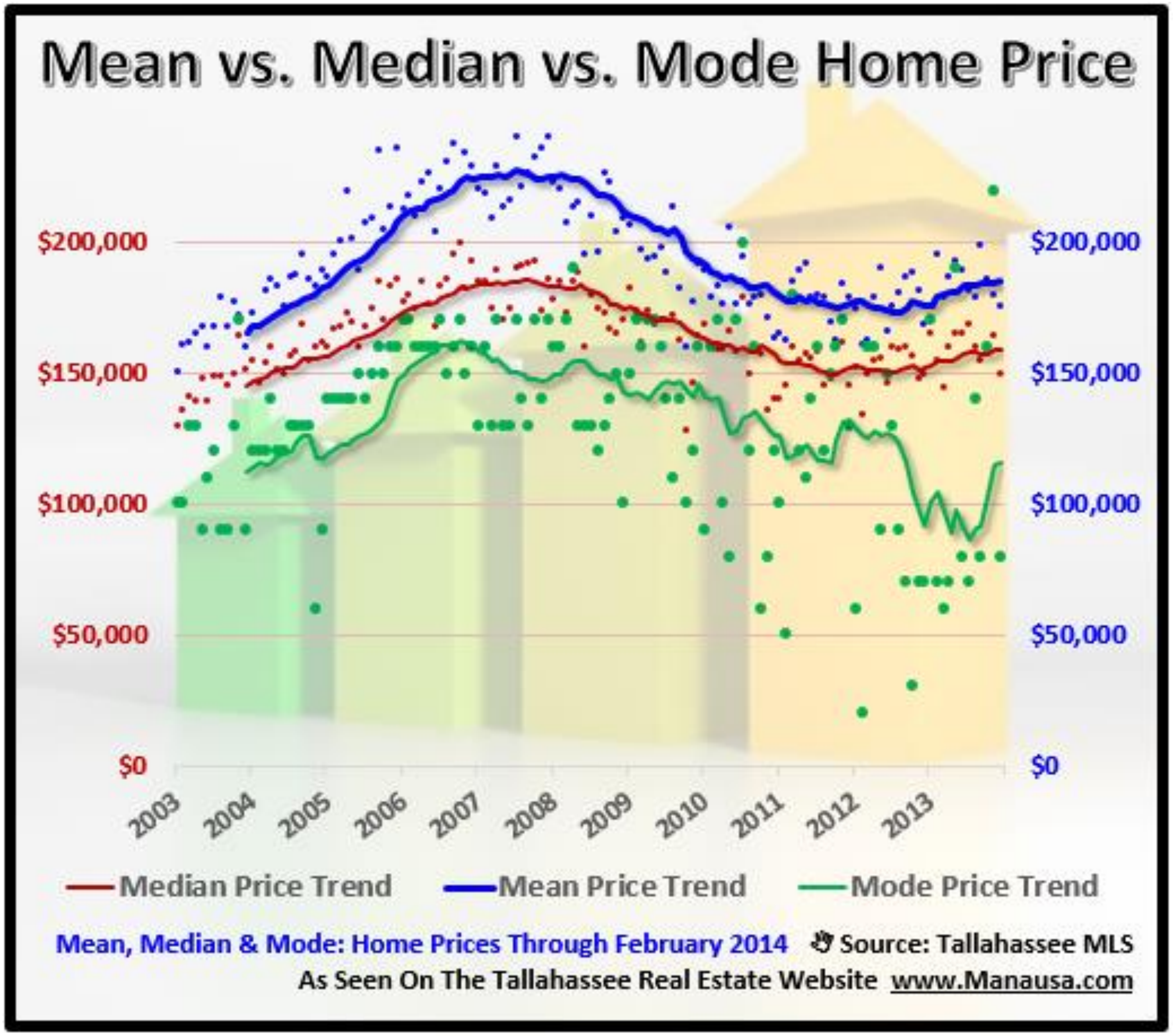
Tuesday, August 2, 2011

Region District	Units	Total Volume	Average Price	6 Month Average	Median Price
Residential					
● Single Family					
Greater Victoria					
Victoria	25	$12,991,300	$519,652	$621,921	$480,000
Victoria West	1	$440,000	$440,000	$443,019	$440,000
Oak Bay	19	$16,384,551	$862,345	$878,931	$740,000
Esquimalt	15	$6,595,000	$439,667	$469,087	$415,000
View Royal	4	$2,607,500	$651,875	$554,075	$696,250
Saanich East	55	$33,451,900	$608,216	$654,985	$590,000
Saanich West	20	$10,344,000	$517,200	$567,056	$532,500
Central Saanich	15	$9,281,400	$618,760	$602,821	$534,900
North Saanich	8	$5,489,900	$686,238	$704,703	$677,500
Sidney	11	$5,321,000	$483,727	$488,759	$480,000
Highlands	1	$643,500	$643,500	$573,229	$643,500
Colwood	13	$6,372,999	$490,231	$503,013	$482,000
Langford	30	$14,815,565	$493,852	$508,104	$473,000
Metchosin	5	$3,028,000	$605,600	$651,776	$603,000
Sooke	18	$7,130,588	$396,144	$408,694	$412,500
Waterfront (all districts)	14	$12,706,400	$907,600	$1,071,544	$880,000
Total Greater Victoria	**254**	**$147,603,603**	**$581,117**	**$615,439**	**$535,000**
Other Areas					
Shawnigan Lake / Malahat	1	$338,000	$338,000	$426,518	$338,000
Gulf Islands	11	$5,434,525	$494,048	$466,101	$440,000
Uplsland / Mainland	12	$5,153,755	$429,480	$453,854	$415,500
Waterfront (all districts)	5	$4,115,000	$823,000	$862,788	$640,000
Total Other Areas	**29**	**$15,041,280**	**$518,665**	**$515,932**	**$450,000**
Total Single Family	**283**	**$162,644,883**	**$574,717**	**$606,341**	**$529,900**



The average price for single-family homes sold in Greater Victoria last month was $581,117, down from $629,292 in June. The median price also declined to $535,000 while the six-month average declined to $615,439. There were 13 single family home sales of over $1 million in July including two on the Gulf Islands. The overall average price for condominiums last month was $315,371, down from $320,172 in June. The average for the last six months declined to $327,762. The median price for condominiums in July also declined to $289,000. The average price of all townhomes sold last month declined to $412,178 from $444,768 in June. The median price also declined to $385,000 while the six month average declined to $443,341.

MLS® sales last month included 283 single family homes, 147 condominiums, 47 townhomes and 19 manufactured homes.



Mean vs. Median vs. Mode Home Price

Mean, Median & Mode: Home Prices Through February 2014 Source: Tallahassee MLS

As Seen On The Tallahassee Real Estate Website www.Manausa.com

(iii) **The M_____**: The “balancing” point of data.
Calculate “mean” or average in several ways:

(i) Simple Arithmetic _____: $\mu = \frac{1}{N}\left[X_1 + X_2 + \cdots + X_{\mathrm{N}}\right] = \frac{1}{N}\sum X_i$

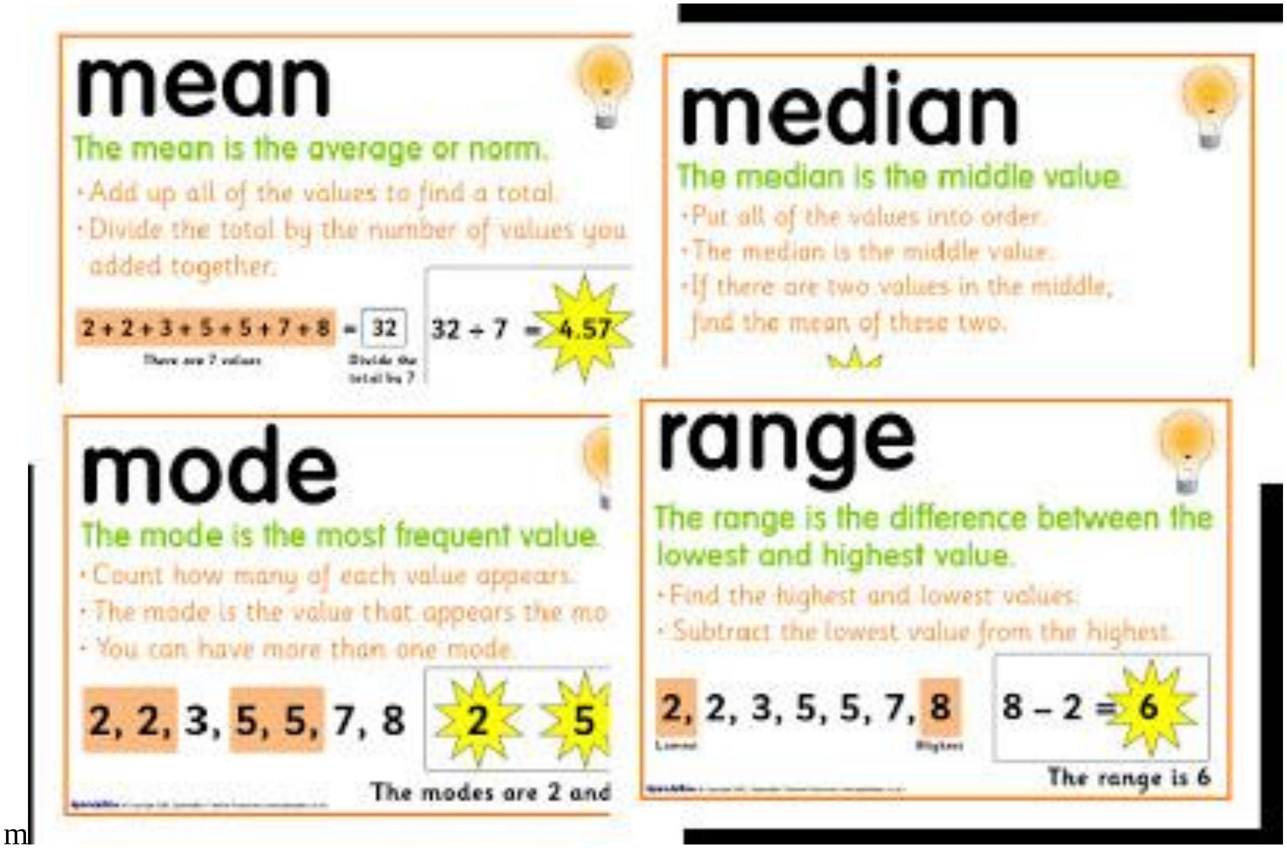
Example: Temperatures: {22, 20, 16, 24, 18, 16, 21, 19, 23} N=9

$$\mu = 19.889$$

Note for comparison:

Putting data is ascending order: 16, 16, 18, 19, 20, 21, 22, 23, 24

↳Mode=16 (occurs twice)
↳Median=20 (5th observation)



mean

The mean is the average or norm.

- Add up all of the values to find a total.
- Divide the total by the number of values you added together.

2 + 2 + 3 + 5 + 5 + 7 + 8 = 32

There are 7 values

Divide the total by 7

32 ÷ 7 = 4.57

median

The median is the middle value.

- Put all of the values into order.
- The median is the middle value.
- If there are two values in the middle, find the mean of these two.

mode

The mode is the most frequent value.

- Count how many of each value appears.
- The mode is the value that appears the mo
- You can have more than one mode.

2, 2, 3, 5, 5, 7, 8

2 5

The modes are 2 and

range

The range is the difference between the lowest and highest value.

- Find the highest and lowest values.
- Subtract the lowest value from the highest.

2, 2, 3, 5, 5, 7, 8

Lowest

Highest

8 – 2 = 6

The range is 6

m

Property: data values “balance” about μ. ***Proof***:

$$\left(X_1 - \mu\right) + \left(X_2 - \mu\right) + \left(X_3 - \mu\right) + \cdots + \left(X_N - \mu\right)$$

$$= \left(X_1 + X_2 + X_3 + \cdots + X_N\right) - \left(\mu + \mu + \mu + \cdots + \mu\right)$$

$$= \sum_{i=1}^{N} X_i - N\mu$$

$$= \left(N\mu - N\mu\right) = 0$$

Since $\frac{\sum_{i=1}^{N} X_i}{N} = \mu$, hence: $\sum_{i=1}^{N} X_i = N\mu$.

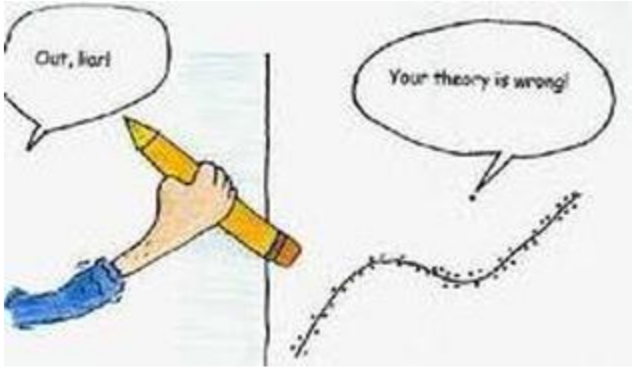
Problems with the Mean:

(A) Affected by __________ in the data.

↳ (One strange observation misrepresents the data.)

(B) Each data point gets _______weight when calculating μ–
⇨unrealistic

↳ (Does not take into account frequency or importance of certain values.)



Variations:

(ii) Weighted Arithmetic ______

In situations in which some values are more important than others, a weighted average should be used.

$$\mu_{\mathrm{W}} = \frac{\left[W_1 X_1 + W_2 X_2 + \cdots + W_N X_N\right]}{\left[W_1 + W_2 + \cdots + W_N\right]}$$

$$\mu_{\mathrm{W}} = \frac{\left[\sum_{i=1}^{N} W_i X_i\right]}{\sum_{i=1}^{N} W_i}$$

Example: Grouped Data

Weight

Income	**Frequency (f_i)**	**Mid-Point (X_i)**
80 $\leq$ X < 100	**2**	
100 $\leq$ X < 120	**6**	
120 $\leq$ X < 140	**8**	
140 $\leq$ X < 160	**6**	
160 $\leq$ X < 180	**3**	
180 $\leq$ X < 200	**13**	
200 $\leq$ X < 220	**2**	
	$\sum f_i = 40$	

If there are <u>K</u> intervals (K=7), then the weighted arithmetic mean is:

$$\mu_W = \frac{\left[\sum_{i=1}^{7} f_i X_i\right]}{\left[\sum_{i=1}^{7} f_i\right]}$$

$$\mu_W = \frac{\left[(90\times 2)+(110\times 6)+\cdots+(210\times 2)\right]}{2+6+8+6+3+13+2}$$

$$\mu_W = 154.6$$

Other "means" are used when the data are all in <u>*ratios*</u> *(or rates of change: i.e. price indices; change in yields on stocks and bonds.)*

(iii) **<u>G________ Mean</u>**: GM gives equal weight to changes of equal relative importance. It cannot be used if any value is ‘0’ in the data.

$$\mu_{\text{G}} = \left(X_1 \times X_2 \times X_3 \times \cdots \times X_N\right)^{1/N} = \left[\prod_{i=1}^{N} X_i\right]^{1/N}$$

where Π is the product operator.

<u>Example: Two interest rates:</u>

$X_1 = 5.4\%$

$X_2 = 6.7\%$

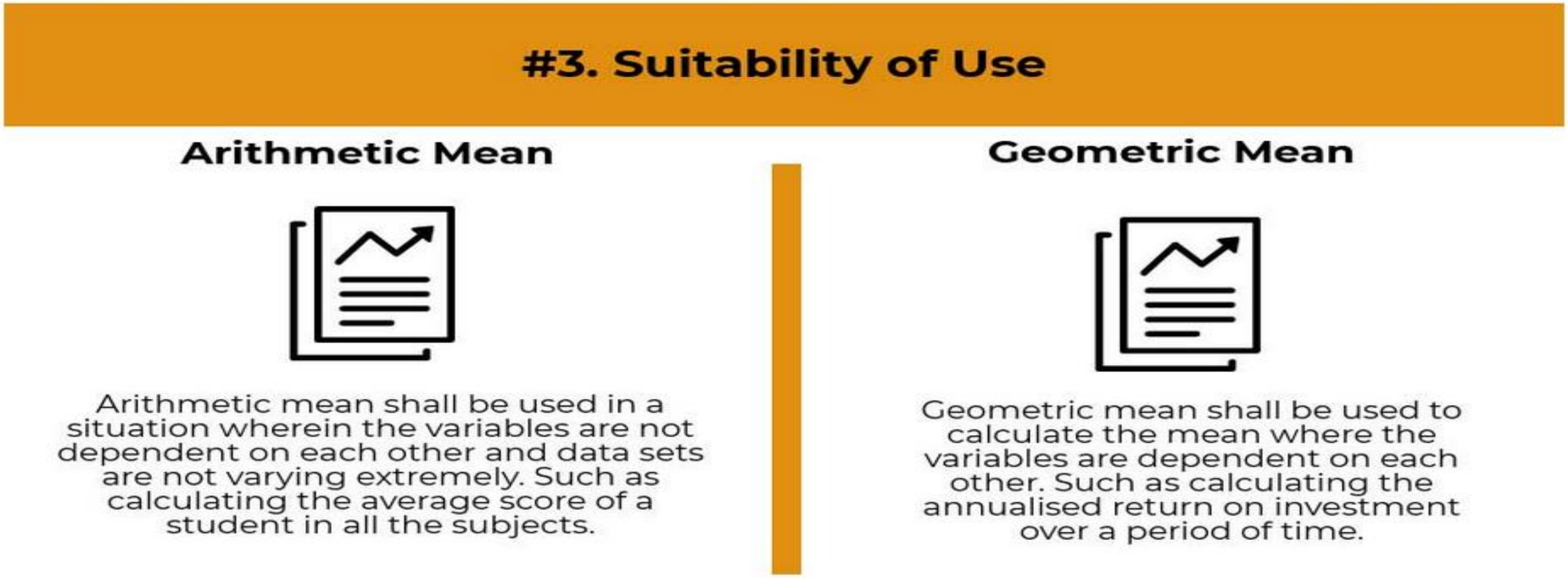
$$\mu_G = (5.4 \times 6.7)^{1/2} = 6.015\%$$

$$\boxed{\mu_G < \mu_A}$$

$$\textit{Note}: \quad \mu = \frac{5.4 + 6.7}{2} = 6.05\%$$

Difference Between Geometric Mean vs Arithmetic Mean

The Arithmetic mean and Geometric mean are the tools widely used to calculate the returns on investment for investment portfolios in the world of finance. People use the arithmetic mean to report the higher returns which are not the correct measure of calculating the return on investment. Since the return on investment for a portfolio over years is dependent on returns in previous years, Geometric mean is the correct way to calculate the return on investment for a specific time period. Arithmetic mean is better suited in the situation wherein variables being used for calculation of average are not dependent on each other.



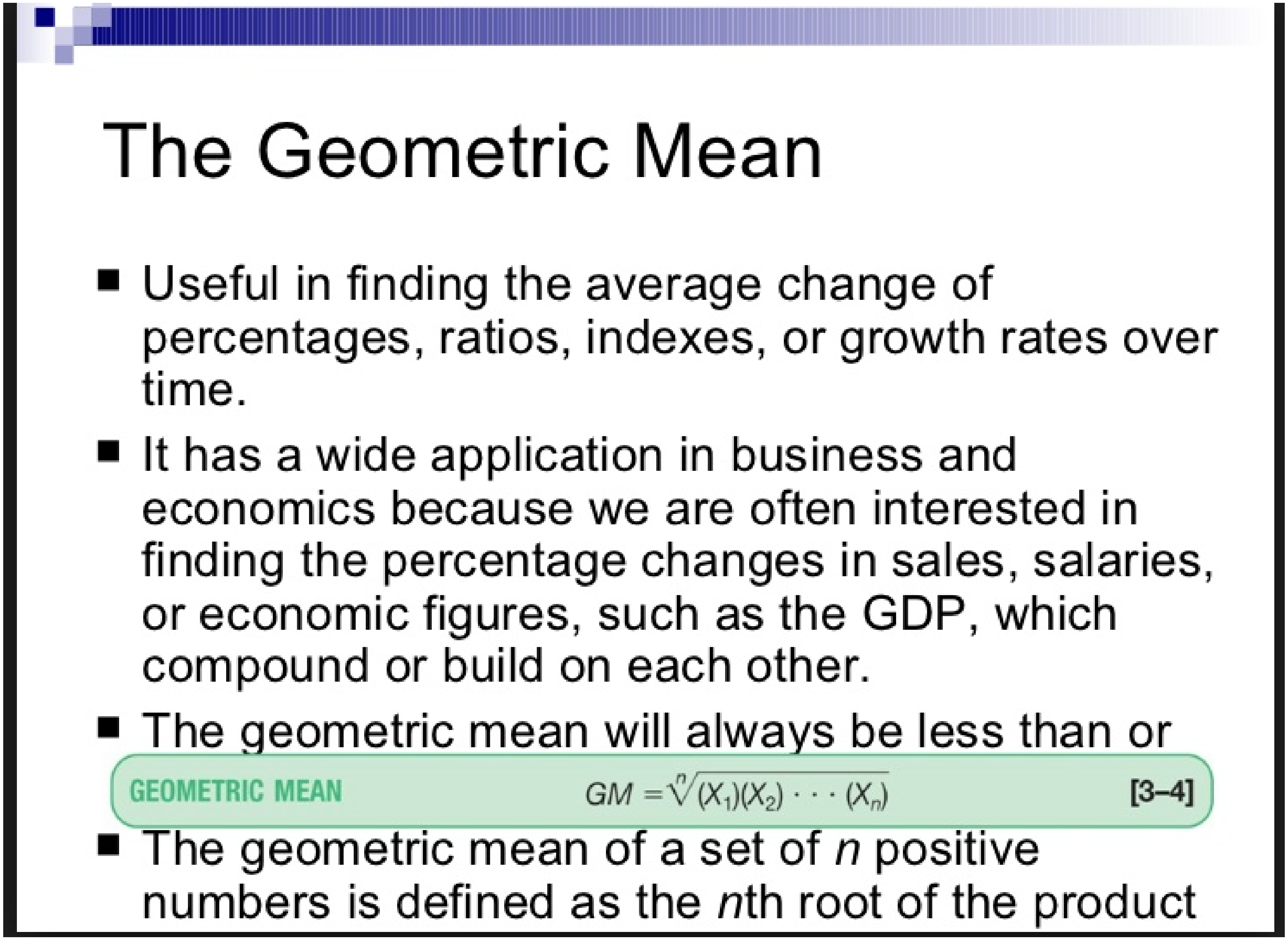
#3. Suitability of Use

Arithmetic Mean

Arithmetic mean shall be used in a situation wherein the variables are not dependent on each other and data sets are not varying extremely. Such as calculating the average score of a student in all the subjects.

Geometric Mean

Geometric mean shall be used to calculate the mean where the variables are dependent on each other. Such as calculating the annualised return on investment over a period of time.



The Geometric Mean

- Useful in finding the average change of percentages, ratios, indexes, or growth rates over time.
- It has a wide application in business and economics because we are often interested in finding the percentage changes in sales, salaries, or economic figures, such as the GDP, which compound or build on each other.
- The geometric mean will always be less than or

GEOMETRIC MEAN $GM = \sqrt[n]{(X_1)(X_2) \cdots (X_n)}$ **[3–4]**

- The geometric mean of a set of *n* positive numbers is defined as the *n*th root of the product

Besides being used by scientists and biologists, geometric means are also used in many other fields, most notably financial reporting. This is because when evaluating investment returns as annual percent change data over several years (or fluctuating interest rates), it is the geometric mean, not the arithmetic mean, that tells you what the average financial rate of return would have had to have been over the entire investment period to achieve the end result. This term is also so called the Compound Annual Growth Rate or CAGR. Population biologists also use the same calculation to determine average growth rates of populations, and this growth rate is referred to as the Intrinsic Rate of Growth when the calculation is applied to estimates of population increases where there are no density-dependent forces regulating the population.

Financial Return Calculation

For financial investment return calculations, the geometric mean is calculated on the decimal multiplier equivalent values, not percent values (i.e., a 6% increase becomes 1.06; a 3% decline is transformed to 0.97. Just follow the steps outlined in the section below titled Calculating Geometric Means with Negative Values).

The equation is also flipped around when calculating the financial rate of return if you know the starting value, end value, and the time period. This equation is used in these cases when the average rate of return is needed (or population growth rate):

$$\text{Return} = \sqrt[\text{Years}]{(\text{Finalvalue/origvalue})}$$

Note: If you subtract 1 from the equation above, this is your compound interest rate. To use this equation, if years=5, this is the "fifth root", which is the same as raising to the power of 1/5 or 0.2).

Things to Note:

(i) $\mu_G < \mu$ if all $X_i > 0$ and **not** all the same.

(ii)

$$\text{Log } \mu_G = Log\left[\prod_{i=1}^{N} X_i\right]^{1/N}$$

$$= \frac{1}{N} Log\left[\prod_{i=1}^{N} X_i\right] = \frac{1}{N} Log(X_1 \times X_2 \cdots \times X_n)$$

$$= \frac{1}{N}(Log\ X_1 + Log\ X_2 + \cdots + Log\ X_N)$$

$$= \frac{1}{N} \sum_{i=1}^{N} Log(X_i)$$

(Arithmetic mean of logs of data.)

In what sense is the geometric mean a _________point for the data?

Recall: $\left(X_1 - \mu\right) + \left(X_2 - \mu\right) + \left(X_3 - \mu\right) + \cdots + \left(X_N - \mu\right) = 0$

In the case of the geometric mean:

$$\left(\frac{X_1}{\mu_G}\right)\left(\frac{X_2}{\mu_G}\right)\left(\frac{X_3}{\mu_G}\right)\cdots\left(\frac{X_N}{\mu_G}\right)$$

$$= \frac{\left(\prod_{i=1}^{N} X_i\right)}{\left(\mu_G\right)^N} = \frac{\left(\mu_G\right)^N}{\left(\mu_G\right)^N} = 1$$

Since $\mu_G = \left[\prod X_i\right]^{1/N}$ *and* $\left[\mu_G = \left[\prod X_i\right]^{1/N}\right]^N \Rightarrow \mu_G{}^N = \left[\prod X_i\right]$

Everything balances multiplicatively about the '_______' ratio of unity – this is why the geometric mean is appropriate with ratio data.

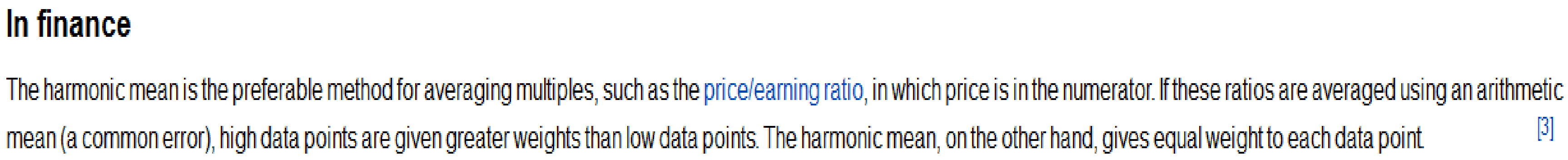
(iv) **H________Mean**: it is the reciprocal of mean of reciprocals.

$$\mu_H = \frac{1}{\left[\frac{1}{N}\sum_{i=1}^{N}\left(\frac{1}{X_i}\right)\right]}$$

Generally: $\mu_H < \mu_G < \mu_A$

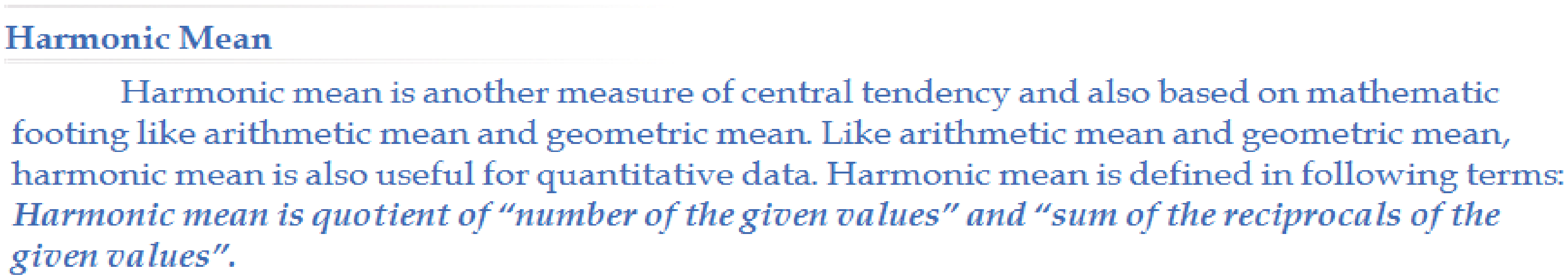
Example: {2, 4, 7, 8}

$$\mu = 5.25$$
$$\mu_G = 4.60$$
$$\mu_H = 3.93$$



In finance

The harmonic mean is the preferable method for averaging multiples, such as the price/earning ratio, in which price is in the numerator. If these ratios are averaged using an arithmetic mean (a common error), high data points are given greater weights than low data points. The harmonic mean, on the other hand, gives equal weight to each data point [3]



Harmonic Mean

Harmonic mean is another measure of central tendency and also based on mathematic footing like arithmetic mean and geometric mean. Like arithmetic mean and geometric mean, harmonic mean is also useful for quantitative data. Harmonic mean is defined in following terms: ***Harmonic mean is quotient of "number of the given values" and "sum of the reciprocals of the given values".***

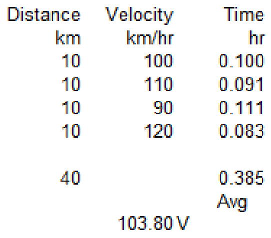
Harmonic Mean

The ***harmonic mean*** is a better "average" when the numbers are defined in relation to some unit. The common example is averaging speed.

For example, suppose that you have four 10 km segments to your automobile trip. You drive your car:

- 100 km/hr for the first 10 km
- 110 km/hr for the second 10 km
- 90 km/hr for the third 10 km
- 120 km/hr for the fourth 10 km.

What is your average speed? Here is a spreadsheet solution:



Distance km	Velocity km/hr	Time hr
10	100	0.100
10	110	0.091
10	90	0.111
10	120	0.083
40		0.385
		Avg
	103.80 V	

The harmonic mean formula is:

$$HM = \frac{n}{\sum_{i=1}^{n} \frac{1}{x_i}} = \frac{4}{\frac{1}{100} + \frac{1}{110} + \frac{1}{90} + \frac{1}{120}} = 103.8$$

Excel calculates this with the formula **`=HARMEAN(100,110,90,120)`**. Unfortunately, the formula is not generalized to average velocities if across different distances.

Harmonic Mean

- **Harmonic mean** (formerly sometimes called the **subcontrary mean**) is one of several kinds of average.
- Typically, it is appropriate for situations when the average of rates is desired. The harmonic mean is the number of variables divided by the sum of the reciprocals of the variables. Useful for ratios such as speed (=distance/time) etc.

What are the applications of harmonic mean?

The harmonic mean is restricted in its field of applications. It is useful for computing the average rate of increase of profits or average speed at which a journey has been performed or the average price at which an article has been sold. For example, if a man walked 20 km., in 5 hours, the rate of his walking speed can be expressed as follows:

$$\frac{20\ km.}{5\ hours} = 4\ km.\ per\ hour,$$

*The general **location** of data is a useful measure, but we need **more**. Extending the idea of the median:*

Per________: *'divide data into 100 equal parts'.*

Rank the data into ascending order.

Method 1:

Calculate positions in the data list below which a _____ percentage of the data lie.

The K^{th} percentile:

$$P_K = \left({(N \times K)}\big/{100} \right)$$

where N = # of observations and K = %.

Then (1) If P_K is integer, add 0.5.

(2) If non-integer, round to next higher integer.

This **locates** the _________ of the K^{th} percentile.

location

Example: (1, 3, 4, 7, 9, 11)   N=6

$P_{30} = \left({(6 \times 30)}/{100} \right) = 1.8$ ⇨ **2nd observation**

The 30th percentile is ____.

(1, 3, 4, 7, 9, 11)

Example:{1, 3, 5, 7, 8, 9, 9, 11 12, 14}   N=10

$$P_{17} = \left({(10 \times 17)} \big/ {100} \right) = 1.7$$ ⇨ **2nd observation**

The 17th percentile is ‘___'.

$$P_{45} = \left({(10 \times 45)} \big/ {100} \right) = 4.5$$ ⇨ 5th **observation**

The 45th percentile is ‘____'.

❄ K

K %

0%

100%

Method 2: Text Method:

$$P_K = \frac{P}{100}(n+1) \leftarrow \text{gives location}$$

Example: (1, 3, 4, 7, 9, 11) N=6

We want the 30th percentile.

$$P_K = \frac{30}{100}(6+1)$$

P_{30} = 0.3(7)=2.1 $\leftarrow$ gives location

The interpretation of P_{30}=2.1 is that 30th percentile is 10%of the way between the value in position 2 and the value is position 3.

The 30th percentile is the value in position 2 plus 0.1 times the difference between the value on position 3 and the value is position 2:

30th percentile =3+0.1(4-3)=3+0.1=3.1

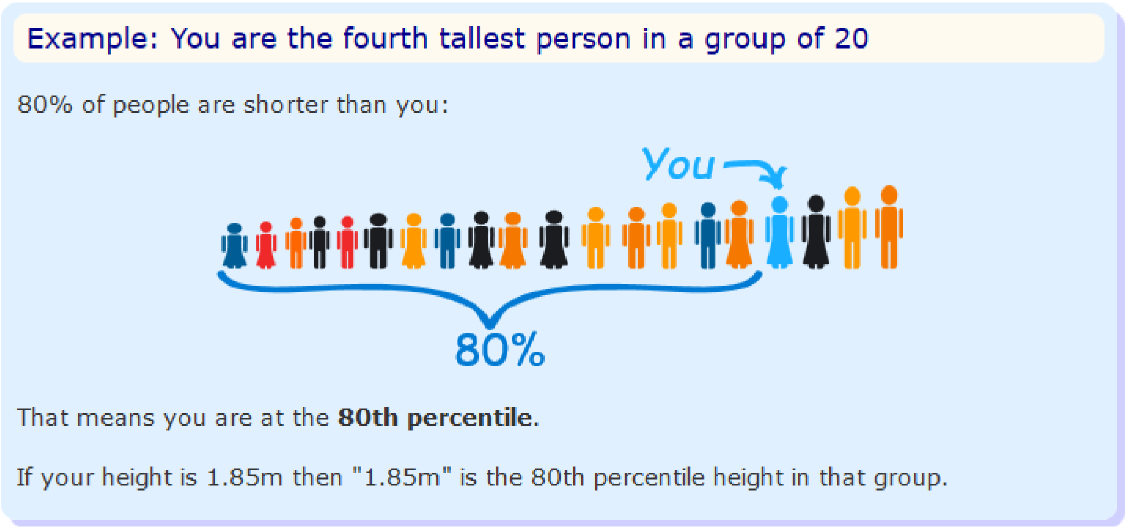
Example:{1, 3, 5, 7, 8, 9, 9, 11 12, 14}   N=10

$$P_{17} = \left({}^{17}\!/_{100} \right)(10+1) = 1.87$$

17th percentile =1+0.87(3-1)=2.74

$$P_{45} = \left({}^{45}\!/_{100} \right)(10+1) = 4.95$$

45th percentile =7+ 0.95(8-7)= 7.9



Example: You are the fourth tallest person in a group of 20

80% of people are shorter than you:

That means you are at the **80th percentile**.

If your height is 1.85m then "1.85m" is the 80th percentile height in that group.

Special Cases:

- ___h Percentile = First Quartile (Q_1)
- 50th Percentile=Second Quartile (Q_2) = Median
- ___th Percentile = Third Quartile (Q_3)
- (Q_3-Q_1) = "Interquartile _________"

Percentiles divide the data into 100 equal parts, each representing one percent of all values.

Eg. The 90th percentile is the value that has 90% of all values below it and 10% above it.

Q_______divide the data into 4 equal parts; Only 3 quartile values are necessary to divide the data into 4 parts.

Example : Interquartile Range

{3, 5, 7, 2, 1}

Step 1: Order data:

{1, 2, 3, 5, 7}   n=5

Step 2: Location

$$Q_3 = P_{75} = \left((5 \times 75) \big/ 100 \right) = 3.75 \quad \Rightarrow 4th$$

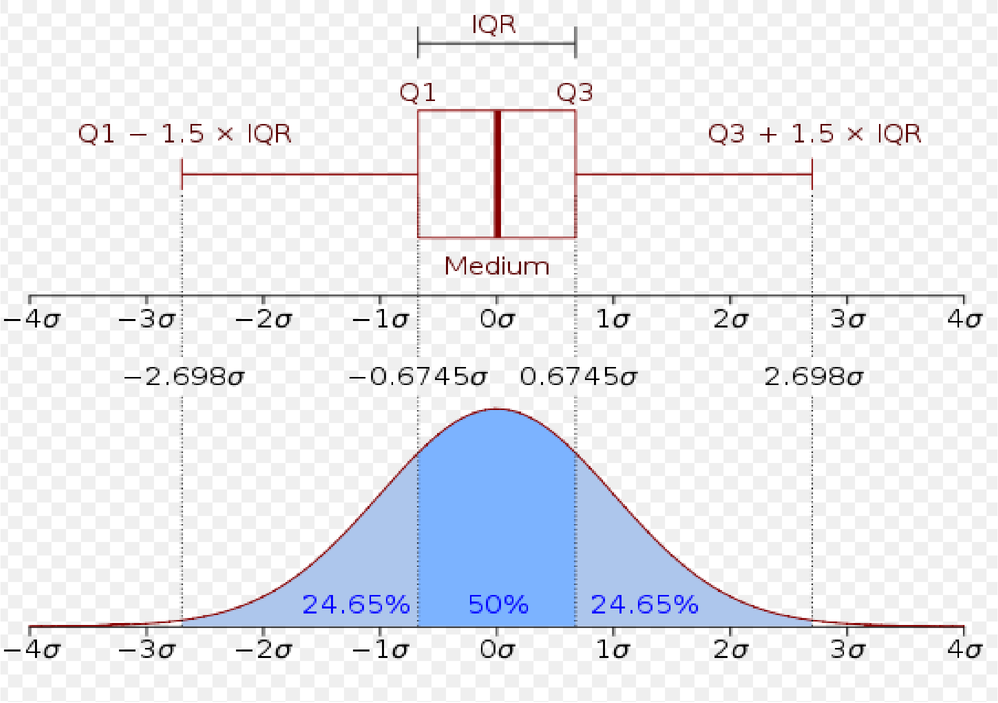
4th observation = 5

$$Q_1 = P_{25} = \left((5 \times 25) \big/ 100 \right) = 1.25 \quad \Rightarrow 2nd$$

2nd observation =2

Step 3: Interquartile calculation:

$$Q_3 - Q_1 = [5 - 2] = 3$$



Unlike (total) range, the interquartile range is a robust statistic, having a breakdown point of 25%, and is thus often preferred to the total range.

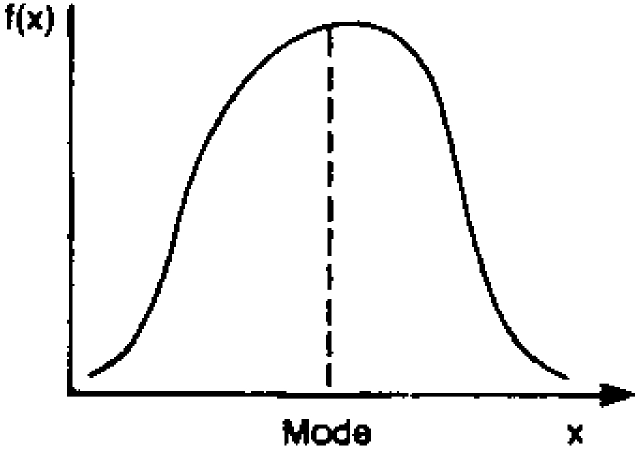
The IQR is used to build box plots, simple graphical representations of a probability distribution.

For a symmetric distribution (so the median equals the midhinge, the average of the first and third quartiles), half the IQR equals the median absolute deviation (MAD).

The median is the corresponding measure of central tendency.

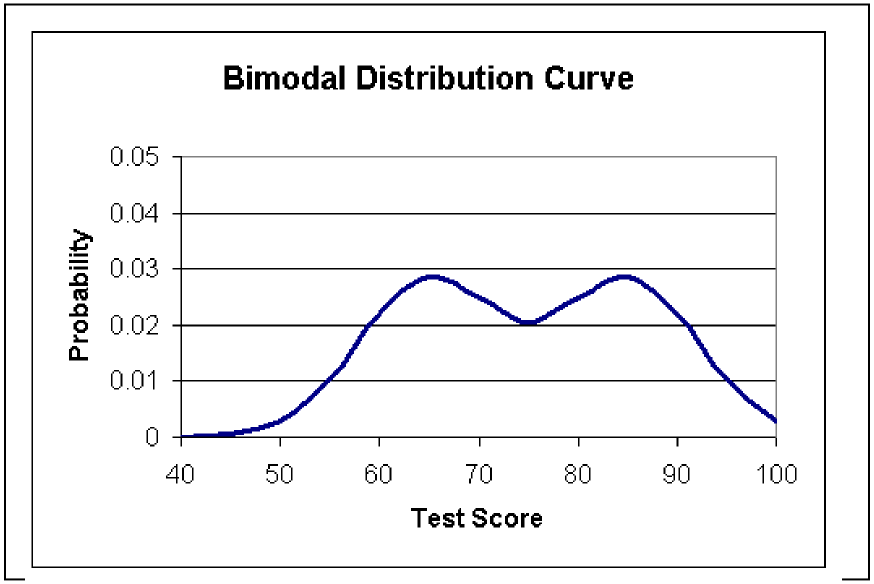
The Shape of A Distribution

● Often having a method for describing the _______ of a frequency distribution is often more helpful than just being able to describe the _________location of a set of data.



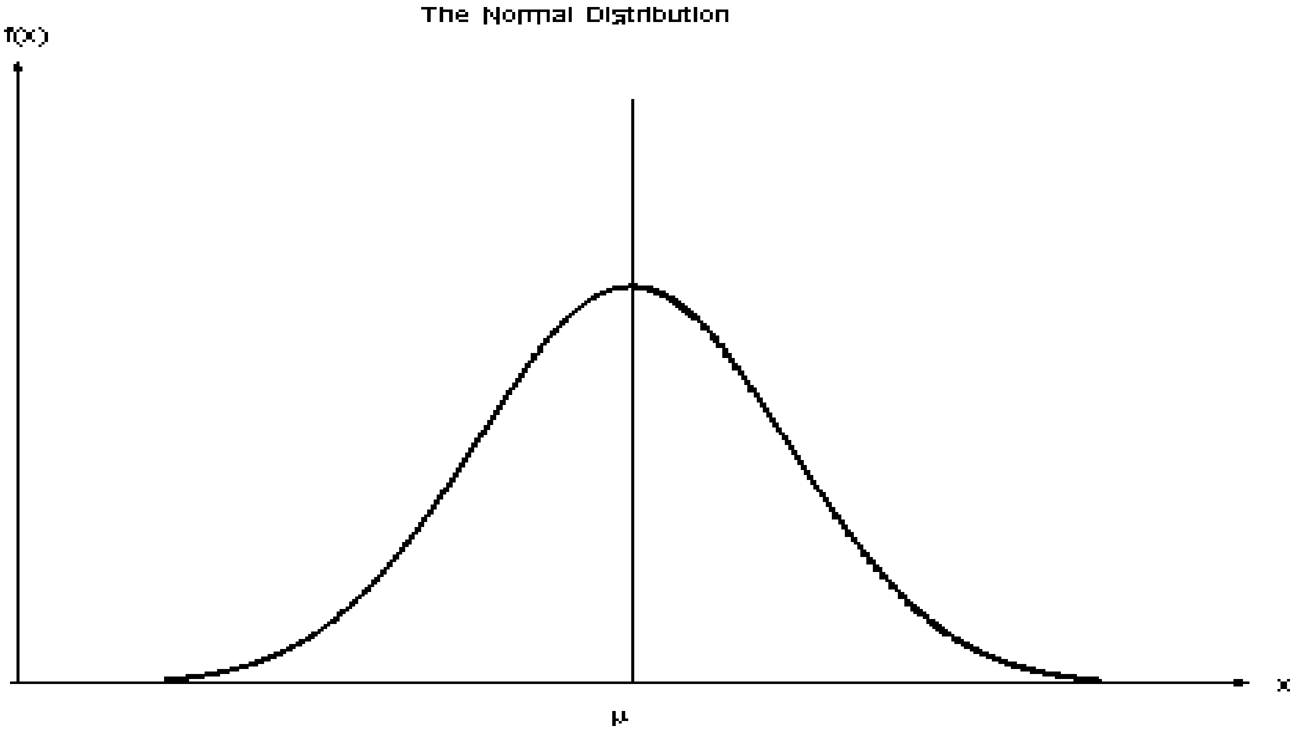
Most "real world" distributions are called _________ distributions, implying one mode or peak.

A distribution with 2 peaks is called a bimodal distribution.



A frequency distribution is said to be **symmetric** if its shape is the same on ____ sides of the median: Median = Arithmetic Mean.
If a distribution is uni-______ and symmetric then:

Mean=Mode=Median



An asymmetric frequency distribution is said to be ________:

1) to the ____ (down) if: mean < median< mode.
2) to the _____(up) if: mean >median> mode.

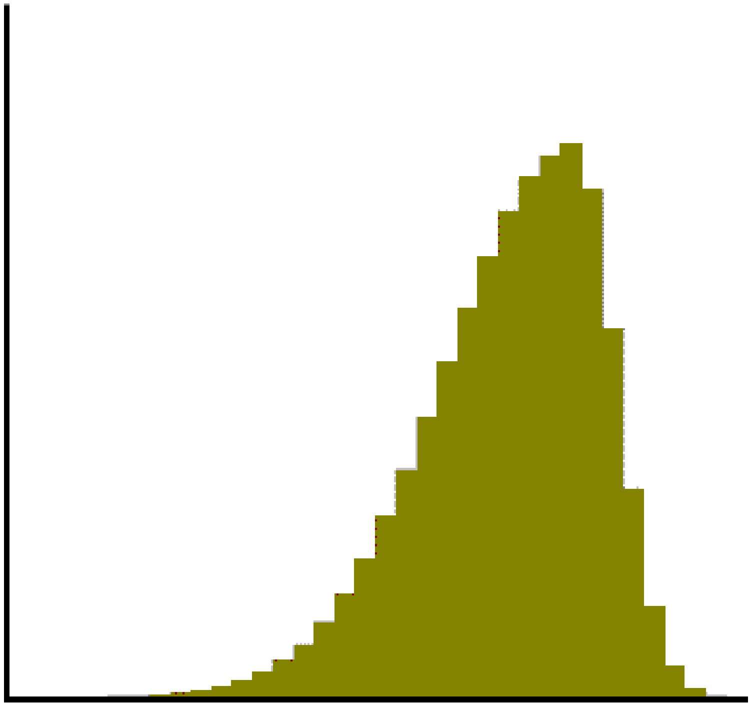


Figure 6.2: Negatively Skewed Distribution

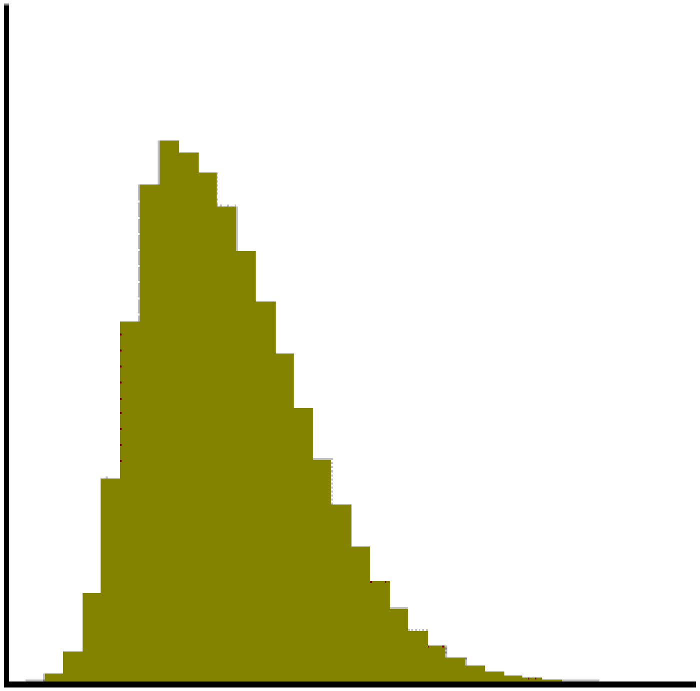


Figure 6.1: Positively Skewed Distribution

Negatively (____) skewed

Positively (_____) skewed

Measures of Dispersion:

●Measures of central location and general ‘shape’ are not enough to properly describe a distribution of data because **variability** or **spread** is __________.

The Range: the range is the _________ difference between the highest and lowest values in the data set.

$R=(X_{max} - X_{min})$

Advantages:

●Independent of measure of _________ location
●Easy to calculate

Disadvantages:

↓Ignores all data except 2 items
↓These values may be "__________." I.e. not representative.

Example: Price of Phones: {10, 37.5, 25, 35, 7, 15, 45, 27}
R=(45-7)=$38

Rank Data:{7 10 15 25 27 35 37.5 45}

(45-7)=38   = RANGE

The Mid-Range(s): Measure of the spread of the innermost concentrated __________ of the data.

- Obtained by excluding a specified proportion of the extreme values at both ends of the ordered values in the data set.

(Order data and exclude a certain proportion at either end of the data set.)

Use these to avoid _________ problems.

Example: Interest Rates

First, order the data:

{6.3, 6.4, 6.5, 7.0, 7.2, 7.5, 8.5, 9.1, 9.2, 11.4} % N=10

80% ___-______ ⇨ discard top and bottom 10% of data (20%)
*Calculate the 10th and 90th percentiles.

$$P_{90} = (90 \times 10)/100 = 9 \text{ add } 0.5 \Rightarrow 9.5 \text{ position:}$$

$$(9.2 + 11.4)/2 = 10.3$$

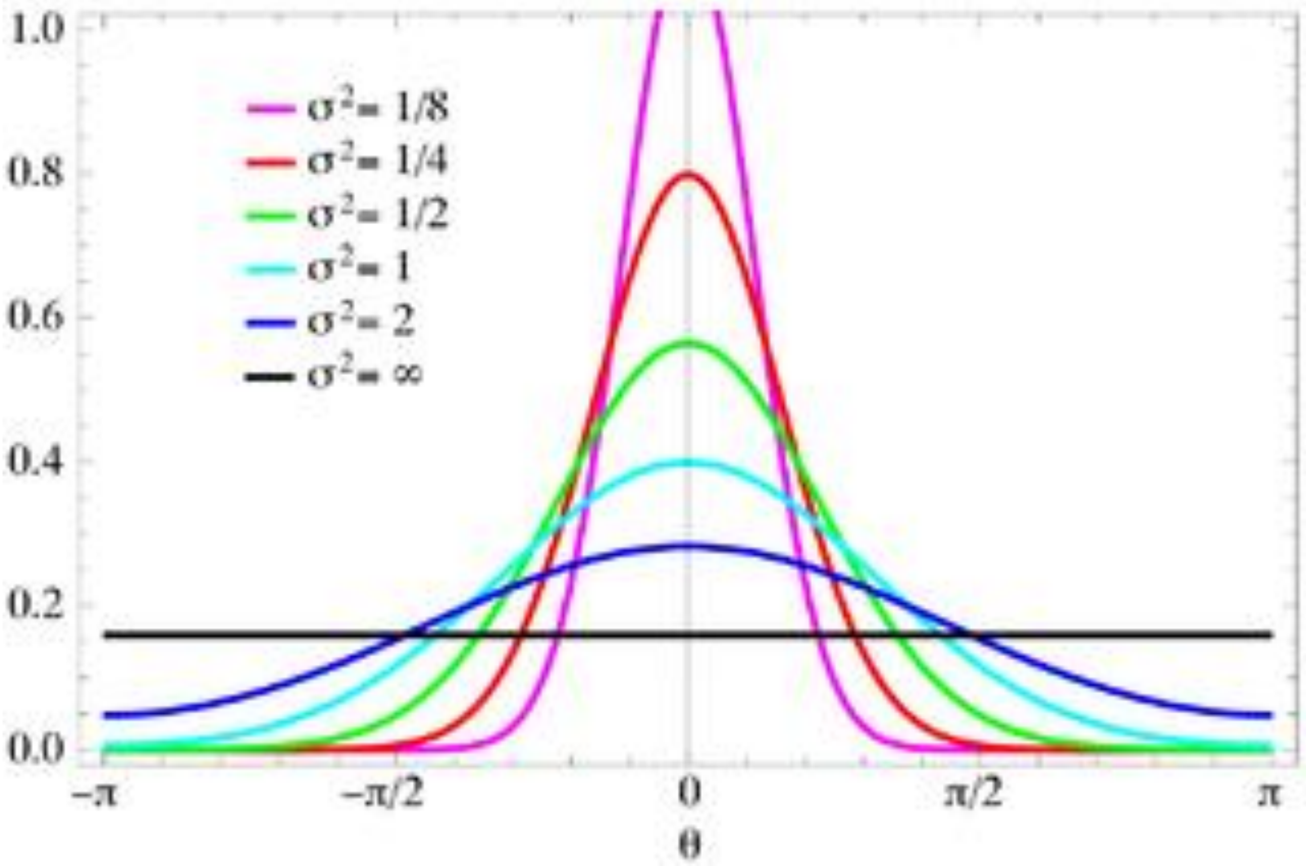
$$P_{10} = (10 \times 10)/100 = 1 \text{ add } 0.5 \Rightarrow 1.5^{th} \text{ position:}$$

$$(6.3 + 6.4)/2 = 6.35$$

R_{80}=(10.3-6.35) =3.95%.

Note: 50% Mid-range= Inter-_________ Range.

A selection of mid-ranges provides a "picture" of the dispersion in data. But, a ______summary measure would be more helpful:



The Variance: A ______ summary measure of data dispersion.

- Takes account of all N data values
- Adjusts the data to take account of ***typical*** level. I.e. use the mean, μ, in calculations.

Construction: Calculate _________from the mean for each observation:

$$\left(X_1 - \mu\right), \left(X_2 - \mu\right), \left(X_3 - \mu\right), \cdots, \left(X_N - \mu\right)$$

We could add these "dispersions" together:

$$\left(X_1 - \mu\right) + \left(X_2 - \mu\right) + \left(X_3 - \mu\right) + \cdots + \left(X_N - \mu\right) = 0$$

But this sum is always _____, and some of the deviations will be positive and some negative:

For example:{1, 2, 3}
↳mean=2

$$(1-2) + (2-2) + (3-2) = (-1) + (0) + (1) = 0$$

So, we need to deal with the positive and negative signs:

***Two Options*:**

Take the "_______" or take "________" values of the individual deviations.

In each case, we need to scale the sum by dividing by N to get the correct order.

This leads to: **Two Measures of Variance**:

(A) **M____ Squared Deviation**: square each deviation and divide by N:

$$MSD = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2$$

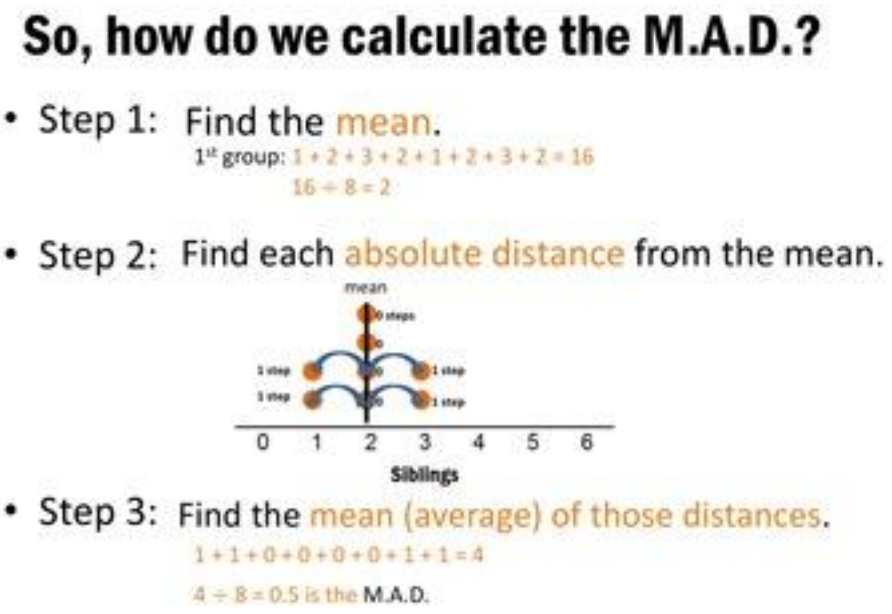
In the case of a population, we also call this the **Population __________**: **(**σ^2 --- sigma squared)

*_**Note the units**_: if data are in $, MSD and the variance are in $^{2}.

So, we often report the standard deviation, which is the square root of the variance: Standard deviation = $\sqrt{\text{Variance}} = \sigma$.

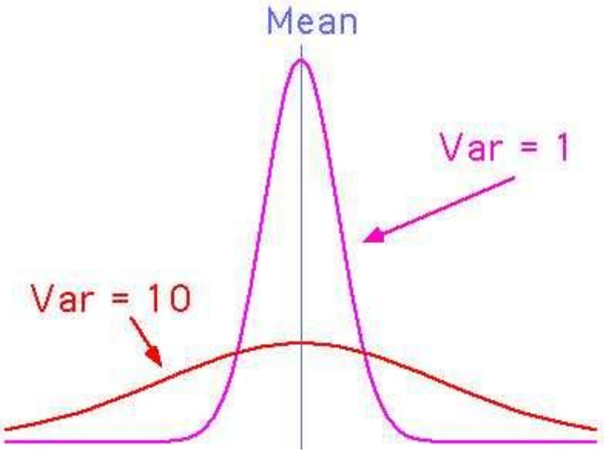
(B) **Mean ___________ Deviation**: Sum the absolute deviation and divide by N:

$$MAD = \frac{1}{N}\sum_{i=1}^{N}\left|X_i - \mu\right|$$



So, how do we calculate the M.A.D.?

- Step 1: Find the mean.
  1st group: 1 + 2 + 3 + 2 + 1 + 2 + 3 + 2 = 16
  16 ÷ 8 = 2
- Step 2: Find each absolute distance from the mean.
- Step 3: Find the mean (average) of those distances.
  1 + 1 + 0 + 0 + 0 + 0 + 1 + 1 = 4
  4 ÷ 8 = 0.5 is the M.A.D.



Example: N=10; Interest Rates; μ=7.91%

i	**X_i**	**$(X_i - \mu)$**	**$(X_i - \mu)^2$**	**$\lvert(Xi-\mu)\rvert$**
1		-1.61	2.5921	1.61
2		-1.51	2.2801	1.51
3		-1.41	1.9881	1.41
4		-0.91	0.8281	0.91
5		-0.71	0.5041	0.71
6		-0.41	0.1681	0.41
7		0.59	0.3481	0.59
8	9.1	1.19	1.4161	1.19
9	9.2	1.29	1.6641	1.29
10	11.4	3.49	12.1801	3.49
		Sum=0	$\sum = 23.969(\%)^2$	$\sum = 13.12(\%)$

Mean square deviation

$$MSD = \sigma^2 = \frac{23.969}{10} = 2.3969(\%)^2$$

$$\sigma = 1.5548\%$$

same _______

$$MAD = \frac{13.12}{10} = 1.312\%$$

Mean absolute deviation

Warning: Rounding error can be an issue.

Alternative Formula for the Variance (MSD):

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i - \mu\right)^2 \Leftarrow \mathit{Expand}$$

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{N}\left(X_i^{\,2} + \mu^2 - 2\mu X_i\right) \Leftarrow \text{Take summation operator through}$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^{\,2} + \underset{\Big\downarrow}{\sum_{i=1}^{N} \mu^2} - 2\mu \underset{\Big\downarrow}{\sum_{i=1}^{N} X_i}\right] \Leftarrow \text{Since } \frac{\sum \mathrm{X_i}}{\mathrm{N}} = \mu$$

$$\text{then: } \mathrm{N}\mu = \sum \mathrm{X_i}$$

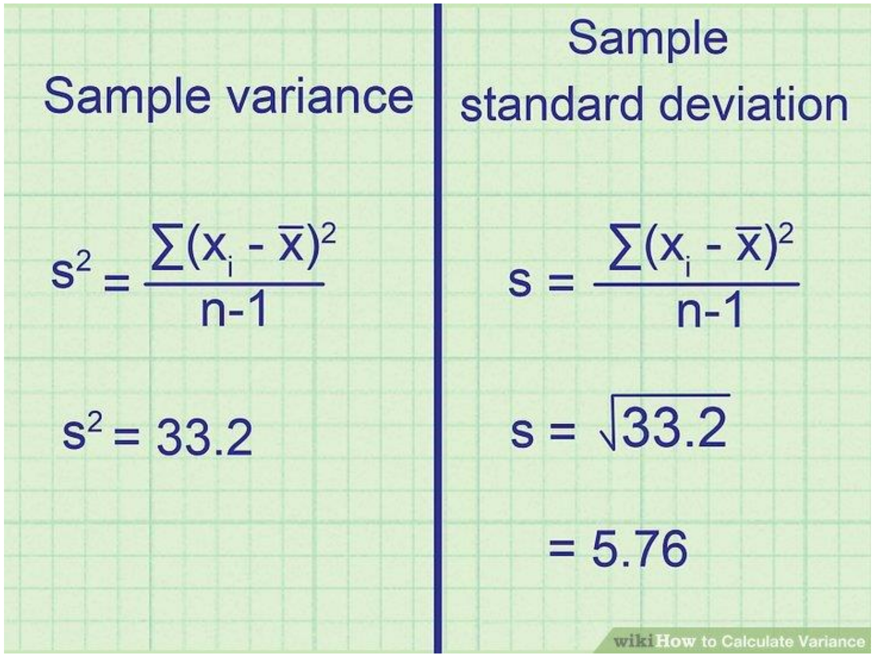
$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^{\,2} + N\mu^2 - 2\mu(N\mu)\right]$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^{\,2} + N\mu^2 - 2N\mu^2\right]$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^{\,2} - N\mu^2\right]$$

$$\sigma^2 = \frac{1}{N}\left[\sum_{i=1}^{N} X_i^{\,2}\right] - \mu^2$$

Note: $$\boxed{\sum_{i=1}^{N}\mu^2 = \mu^2 + \mu^2 + \mu^2 + \cdots + \mu^2 = N\mu^2}$$



Sample variance

$$s^2 = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$s^2 = 33.2$$

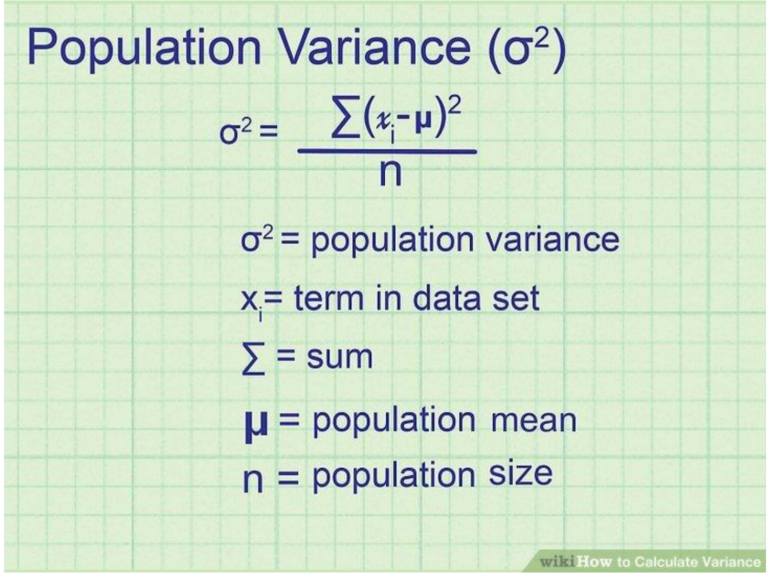
Sample standard deviation

$$s = \frac{\sum(x_i - \bar{x})^2}{n-1}$$

$$s = \sqrt{33.2}$$

$$= 5.76$$

wikiHow to Calculate Variance



Population Variance (σ^2)

$$\sigma^2 = \frac{\sum(x_i - \mu)^2}{n}$$

σ^2 = population variance

x_i = term in data set

$\sum$ = sum

μ = population mean

n = population size

wikiHow to Calculate Variance

What About Grouped Data?

For the Variance:

$$\sigma^2 = \frac{1}{N}\sum_{i=1}^{K}\left[\left(X_i - \mu\right)^2 f_i\right]$$

where:

K =# of groups

f_i = frequency $\Rightarrow \sum f_i = N$

For the Mean Absolute Deviation:

$$MAD = \frac{1}{N}\sum_{i=1}^{K}\left[\left|X_i - \mu\right| f_i\right]$$

$NOTE\text{:}\ X_i$ = midpoint

Example: House Prices ($'000): calculate the variance and SD.

Class (i)	Mid-Point (X_i)	Frequency (f_i)	$(Xi-\mu)^2$	$(Xi-\mu)^2 f_i$
$100<X\leq 200$		3	20,736	62,208
$200<X\leq 300$		27	1,936	52,272
$300<X\leq 400$		15	3,136	47,040
$400<X\leq 500$		5	24,336	121,680
		N=50		283,200

$$\mu = \frac{1}{N}\sum X_i f_i$$

$$\mu = \frac{1}{50}\left[(150\times 3)+(250\times 27)+(350\times 15)+(450\times 5)\right]$$

$$\mu = 294$$

$$\sigma^2 = \frac{1}{50}\left(283{,}200\right) = 5664\ (\$'000)^2$$

$$\sigma = 75.26 \quad (\$'000)$$

Issue: Accuracy With Grouped Data

- We used **midpoints** of groups as representative values for the calculations of mean and ___________.
  - The effect on the _______ is negligible.
- The effect on the variance is **not**! The variance is __________!

Must use **S ____________'s Correction** (for the variance):

$$\sigma_C^2 = \sigma^2 - \left(h^2 \big/ 12 \right)$$

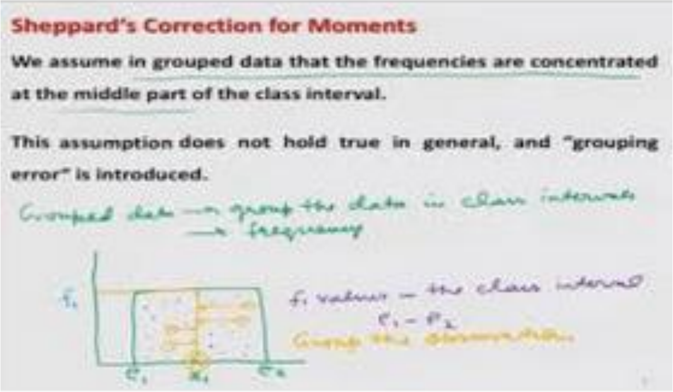
where h = class width

$$\sigma^2 = 5664 \quad (\$'000)^2$$

Example: $\sigma_C^2 = 4830.7 \quad (\$'000)^2$

- Using intervals instead of actual data values, the variance is **over__________**.

- As the interval width gets smaller and smaller, the overestimate ___________. When using individual values, there is no reason to use Sheppard's Correction.



Sheppard's Correction for Moments

We assume in grouped data that the frequencies are concentrated at the middle part of the class interval.

This assumption does not hold true in general, and "grouping error" is introduced.

Comparing Dispersions of Two Populations

How do you compare __________ deviations if:

(i) Scale (average value) differs and/or

(ii) Units of measurement differ across populations? I.e. £,¢,$.

☐**Must construct a _______ measure which is scale free:**

Coefficient of Variation (CV)

$$C.V. = \left[\frac{\sigma}{\mu} \times 100 \right] \%$$

Example:

Population 1 ($)	**Population 2 (Kg.)**
2	
1	
3	
4	

$$\mu = 2.5$$
$$\sigma = 1.118(\$)$$
$$C.V. = 44.72\%$$

$$\mu = 104.33$$
$$\sigma = 2.055(kg.)$$
$$C.V. = 1.97\%$$

Measuring Skewness

Recall:

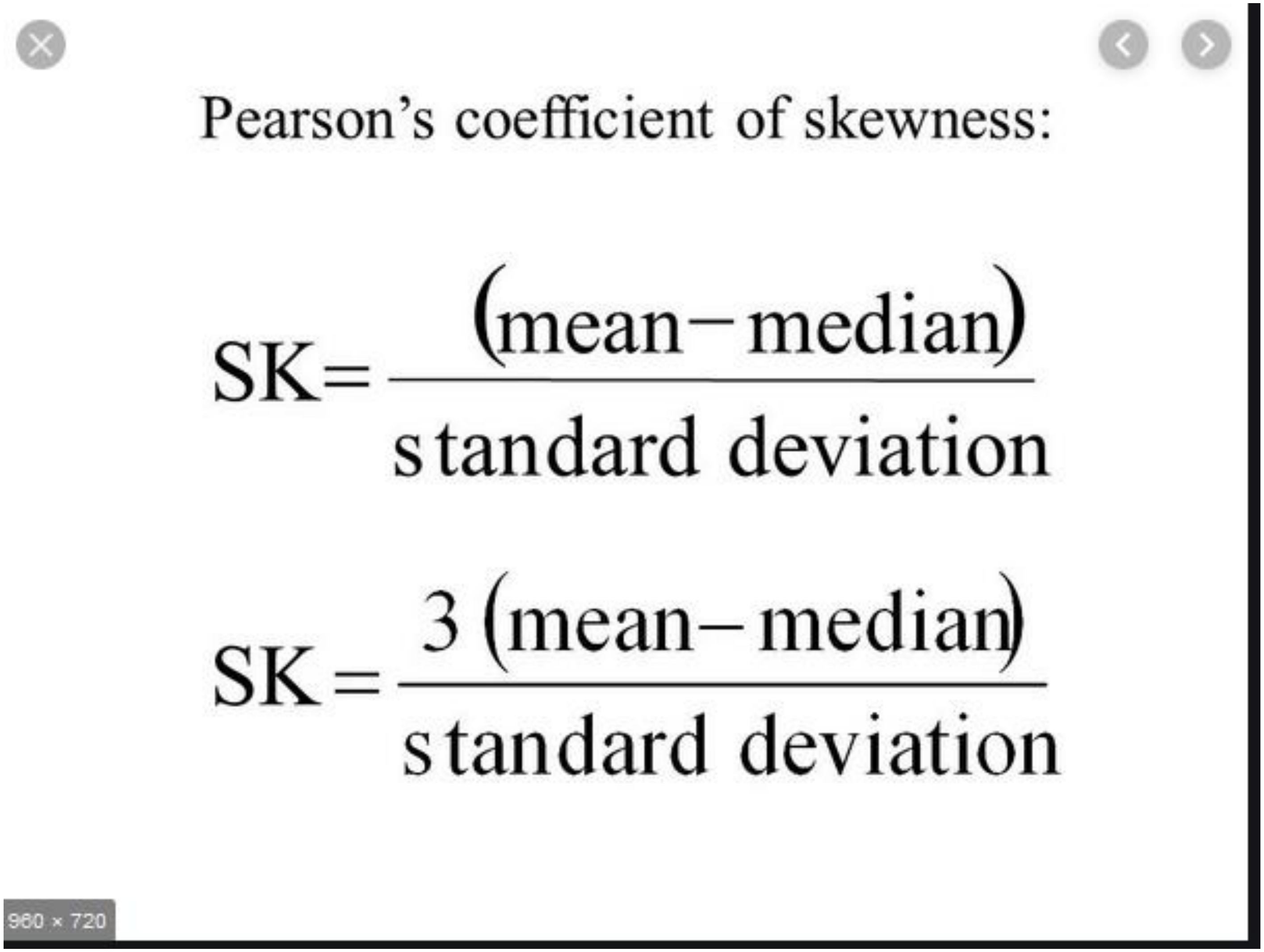
(i) Skewed negatively if mean _____ median.
(ii) Skewed positively if mean _____ median.

Both measures have the same _____.

Pearson's Skewness Measure:

$$\text{Skew} = \left[\frac{\text{mean} - \text{median}}{\sigma}\right] \Rightarrow \text{"+"} \quad \text{"0"} \quad \text{"−"}$$

Unitless measure for comparison. (Range from -1 to 1)



Pearson's coefficient of skewness:

$$SK = \frac{(\text{mean} - \text{median})}{\text{standard deviation}}$$

$$SK = \frac{3\,(\text{mean} - \text{median})}{\text{standard deviation}}$$

Example: {25, 22, 31, 35, 30, 27, 28, 45, 50, 100}

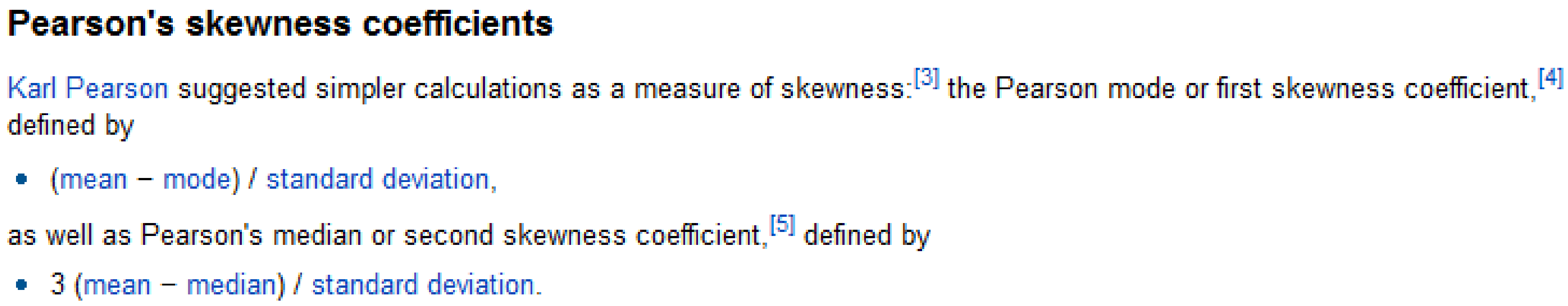
$$\mu = 39.3$$

$$median = 30.5$$

$$\sigma = 21.882$$

$$skew = \left[\frac{39.3 - 30.5}{21.882} \right] = 0.402$$

Unitless –O.K. for comparisons.



Pearson's skewness coefficients

Karl Pearson suggested simpler calculations as a measure of skewness:[3] the Pearson mode or first skewness coefficient,[4] defined by

- (mean − mode) / standard deviation,

as well as Pearson's median or second skewness coefficient,[5] defined by

- 3 (mean − median) / standard deviation.